Vol. XXXIII. Part III November, 1945

BIOMETRIKA

A JOURNAL FOR THE STATISTICAL STUDY OF BIOLOGICAL PROBLEMS

FOUNDED BY

W. F. R. WELDON, FRANCIS GALTON AND KARL PEARSON

EDITED BY

KARL PEARSON

ASSISTED BY

EGON S. PEARSON

Reprinted by offset-litho

ISSUED BY THE BIOMETRIC LABORATORY
UNIVERSITY COLLEGE, LONDON
AND PRINTED AT THE
UNIVERSITY PRESS, CAMBRIDGE

PRINTED IN GREAT BRITAIN

VOLUME XXXIII, PART III NOVEMBER 1945

ON THE USE OF MATRICES IN CERTAIN POPULATION MATHEMATICS

BY P. H. LESLIE, *Bureau of Animal Population, Oxford University*

CONTENTS

	PAGE
1. Introduction	183
2. Derivation of the matrix elements	184
3. Numerical example	185
4. Properties of the basic matrix	187
5. Transformation of the co-ordinate system	188
6. Rèlation between the canonical form B and the $L_x m_x$ column	190
7. The stable age distribution	191
8. Properties of the stable vectors	192
9. The spectral set of operators	193
10. Reduction of B to classical canonical form	194
11. The relation between ϕ and ψ vectors	195
12. Case of repeated latent roots	197
13. The approach to the stable age distribution	199
14. Special case of the matrix with only a single non-zero F_x element	200
15. Numerical comparison with the usual methods of computation	201
16. Further practical applications	207
Appendix: (1) The tables of mortality and fertility	209
(2) Calculation of the rate of increase	210
(3) Numerical values of the matrix elements	212
References	212

1. INTRODUCTION

If we are given the age distribution of a population on a certain date, we may require to know the age distribution of the survivors and descendants of the original population at successive intervals of time, supposing that these individuals are subject to some given age-specific rates of fertility and mortality. In order to simplify the problem as much as possible, it will be assumed that the age-specific rates remain constant over a period of time, and the female population alone will be considered. The initial age distribution may be entirely arbitrary; thus, for instance, it might consist of a group of females confined to only one of the age classes.

The method of computing the female population in one unit's time, given any arbitrary age distribution at time t, may be expressed in the form of $m+1$ linear equations, where m to $m+1$ is the last age group considered in the complete life table distribution, and when the same unit of age is adopted as that of time. If

n_{xt} = the number of females alive in the age group x to $x+1$ at time t,

P_x = the probability that a female aged x to $x+1$ at time t will be alive in the age group $x+1$ to $x+2$ at time $t+1$,

F_x = the number of daughters born in the interval t to $t+1$ per female alive aged x to $x+1$ at time t, who will be alive in the age group 0–1 at time $t+1$,

then, working from an origin of time, the age distribution at the end of one unit's interval will be given by

$$\begin{aligned} \sum_{x=0}^{m} F_x n_{x0} &= n_{01} \\ P_0 n_{00} &= n_{11} \\ P_1 n_{10} &= n_{21} \\ P_2 n_{20} &= n_{31} \\ &\vdots \quad\quad \vdots \\ P_{m-1} n_{m-1,0} &= n_{m1} \end{aligned}$$

or, employing matrix notation, $Mn_0 = n_1$, where n_0 and n_1 are column vectors giving the age distribution at $t = 0$ and 1 respectively, and the matrix

$$M = \begin{bmatrix} F_0 & F_1 & F_2 & \dots & . & F_k & F_{k+1} & \dots & F_{m-1} & F_m \\ P_0 & . & . & \dots & . & . & . & \dots & . & . \\ . & P_1 & . & \dots & . & . & . & \dots & . & . \\ . & . & P_2 & \dots & . & . & . & \dots & . & . \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ . & . & . & \dots & P_{k-1} & . & . & \dots & . & . \\ . & . & . & \dots & . & P_k & . & \dots & . & . \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ . & . & . & \dots & . & . & . & \dots & P_{m-1} & . \end{bmatrix} \quad 0 < P_x < 1;\ F_x \geqq 0.$$

This matrix is square and consists of $m+1$ rows and $m+1$ columns. All the elements are zero, except those in the first row and in the subdiagonal immediately below the principal diagonal. The P_x figures all lie between 0 and 1, while the F_x figures are by definition necessarily positive quantities. Some of the latter, however, may be zero, their number and position depending on the reproductive biology of the species we happen to be considering in any particular case, and on the relative span of the pre- and post-reproductive ages. If $F_m = 0$, the matrix M is singular, since the determinant $| M | = 0$.

Since $Mn_0 = n_1$, and $Mn_1 = M^2_{n_0} = n_2$, etc., the age distribution at time t may be found by pre-multiplying the column vector $\{n_{00} n_{10} n_{20} \dots n_{m0}\}$, i.e. the age distribution at $t = 0$, by the matrix M^t. Moreover, it will be seen that with the help of the jth column of M^t the age distribution and number of the survivors and descendants of the $n_{j-1,0}$ individuals, who were alive at $t = 0$, can readily be calculated. Thus, $n_{j-1,0}$ times the sum of the elements in the jth column of M^t gives the number of living individuals contributed to the total population at time t by this particular age group.

2. Derivation of the matrix elements

The basic data, from which the numerical elements of this matrix may be derived, are given usually in the form of a life table and a table of age specific fertility rates. To take the P_x figures first; if at $t = 0$ there are n_{x0} females alive in the age group x to $x+1$, the survivors of these will form the $x+1$ to $x+2$ age group in one unit's time, and thus $P_x n_{x0} = n_{x+1,1}$. Then it is usually assumed (e.g. Charles, 1938, p. 79; Glass, 1940, p. 464) that

$$P_x = \frac{L_{x+1}}{L_x},$$

where

$$L_x = \int_x^{x+1} l_x \, dx,$$

or the number alive in the age group x to $x+1$ in the stationary or life table age distribution. This method of computing the survivors in one unit's time would be exact if the distribution of those alive within a particular age group was the same as in the life-table distribution.

The F_x figures are more troublesome, and in the numerical example which will be given later they were obtained from the basic maternal frequency figures (m_x = the number of live daughters born per unit of time to a female aged x to $x+1$) by an argument which ran as follows. Consider the n_{x0} females alive at $t = 0$ in the age group x to $x+1$, and let us sup-

pose that they are concentrated at the midpoint of the group, $x+\frac{1}{2}$. During the interval of time 0–1 some of these individuals are dying off, and at $t=1$ the $n_{x+1,1}$ survivors can be regarded as concentrated at the age $x+1\frac{1}{2}$. Although these deaths are taking place continuously, we may assume them all to occur around $t=\frac{1}{2}$, so that at this latter time the number of females alive in the age group we are considering changes abruptly from n_{x0} to $n_{x+1,1}=P_x n_{x0}$. Then during the time interval 0–$\frac{1}{2}$ these n_{x0} females will have been exposed to the risk of bearing daughters, and the number of the latter they will have given birth to per female alive will be given by the maternal frequency figure for the ages $x+\frac{1}{2}$ to $x+1$. This figure may be obtained by interpolating in the integral curve of the m_x values, and thus expressing the latter in $\frac{1}{2}$ units of age throughout the reproductive span instead of in single units. The daughters born during the interval of time 0–$\frac{1}{2}$ will be aged $\frac{1}{2}$–1 at $t=1$, the number of them surviving at this time being determined approximately by multiplying the appropriate $m_{x+\frac{1}{2}}$ figure by the factor $2\int_{\frac{1}{2}}^{1} l_x dx$ according to the given life table. Similarly, each of the $P_x n_{x0}$ females during the interval of time $\frac{1}{2}$–1 give birth to $m_{x+1-x+1\frac{1}{2}}$ daughters, the survivors of which form part of the 0–$\frac{1}{2}$ age group at $t=1$. The survivorship factor is in this case taken to be $2\int_{0}^{\frac{1}{2}} l_x dx$.

Combining these two steps together we obtain a series of F_x figures, which may be defined as the number of daughters alive in the age group 0–1 at $t=1$ per female alive in the age group x to $x+1$ at $t=0$. Putting

$$k_1 = 2\int_{0}^{\frac{1}{2}} l_x dx, \quad k_2 = 2\int_{\frac{1}{2}}^{1} l_x dx,$$

then

$$F_x = (k_2 m_{x+\frac{1}{2}-x+1} + k_1 P_x m_{x+1-x+1\frac{1}{2}}),$$

and

$$\sum_{x=0}^{m} F_x n_{x0} = n_{01},$$

the total number of daughters alive aged 0–1 at $t=1$.

3. NUMERICAL EXAMPLE

In order to see whether the P_x and F_x figures obtained in this way from the basic data give a reasonably accurate estimate of the population in one unit's time, a numerical example was worked out for an imaginary rodent population, the species chosen being the brown rat, *Rattus norvegicus*. Full details of the basic life table and fertility table which were used are given in an appendix, together with a short account of the genesis of these tables and the methods employed to estimate the rate of natural increase (r) and the stable age distribution. Compared with man, the fertility of this imaginary rat population was relatively very great; thus, the gross reproduction rate was 31·21 daughters and the net rate (R_0) 25·66, the life table used being a reasonably good one. The inherent rate of natural increase was estimated to be 0·44565 per head per month of 30 days, and the stable age distribution was so overladen with young that the proportion of females in the post-reproductive age groups was negligible. Some 74·45 % of the females were younger than 3 months, at which age breeding was assumed to commence.

By definition the Malthusian age distribution is stable; that is to say, once a population subject to the given rates of fertility and mortality achieves this form of distribution, it

continues to increase e^r times every unit of time and the proportions of the population alive in each group remain constant. Thus, in the present example, given 100,000 females distributed as to age in the stable form at $t = 0$, the number alive in each age group in 1 month's time can be immediately calculated by multiplying each element in the original distribution by 1·561505. This 'true' age distribution at $t = 1$ is compared in Table 1 with that obtained by operating on the original distribution with the P_x and F_x figures, which are given in Table 5 of the Appendix.

The agreement between the true and estimated age distributions is remarkably close. It might be expected that the principal errors would occur in the early age groups, since the

Table 1

(1) (Units of 30 days) Age group	(2) Population at $t=0$ Stable age distribution	(3) Expected population at $t=1$ Col. 2 × 1·561505	(4) Population at $t=1$ Estimated by operating on col. 2 with the matrix M
0–	37,440	58,463	58,374
1–	22,595	35,282	35,455
2–	14,417	22,512	22,519
3–	9,227	14,408	14,406
4–	5,903	9,218	9,218
5–	3,775	5,895	5,895
6–	2,413	3,768	3,768
7–	1,542	2,408	2,407
8–	984	1,537	1,537
9–	627	979	980
10–	399	623	623
11–	254	397	396
12–	161	251	251
13–	101	158	159
14–	64	100	99
15–	40	62	62
16–	25	39	39
17–	15	23	24
18–	9	14	14
19–	6	9	8
20–	3	5	5
Total	100,000	156,151	156,239

The span of the reproductive ages is from 3 to 21 months.

P_x figures are based on the stationary age distribution which is clearly very different from the stable form. However, as will be seen from Table 1, the biggest error from this cause is due to the first P_0 which overestimates the number alive in the 1–2 age group at $t = 1$ by some 0·5 %. The F_x figures underestimate the number alive in the 0–1 group by 0·2 %, and the total population is overestimated by 0·06 %. On the whole these results are satisfactory and, judging from this example, it would seem that the matrix M operating on a given age distribution should give a reasonable estimate of the population in one unit's time, provided that the unit of time and age chosen be not too coarse as compared with the life span of the species. The degree of cumulative error which is introduced by continued operation with the matrix will be considered later.

4. PROPERTIES OF THE BASIC MATRIX

The matrix M is square and of order $m+1$; it is not necessary, however, in what follows to consider this matrix as a whole. For, if $x = k$ is the last age group within which reproduction occurs, F_k is the last F_x figure which is not equal to zero. Then, if the matrix be partitioned symmetrically at this point,

$$M = \begin{bmatrix} A & . \\ B & C \end{bmatrix}.$$

The submatrix A is square; B is of order $(m-k)\times(k+1)$; C again is square consisting of $m-k$ rows and columns, the only numerical elements being in the subdiagonal immediately below the principal diagonal. The remaining submatrix is of order $(k+1)\times(m-k)$ and consists only of zero elements. Then in forming the series of matrices M^2, M^3, M^4, etc.,

$$M^t = \begin{bmatrix} A^t & . \\ f(ABC) & C^t \end{bmatrix}.$$

The submatrix C is, however, of such a type that $C^{m-k} = 0$, so that M^t, $t \geqq m-k$, will have all its last $m-k$ columns consisting of zero elements. This is merely an expression of the obvious fact that individuals alive in the post-reproductive ages contribute nothing to the population after they themselves are dead. It is the submatrix A which is principally of interest, and in the mathematical discussion which follows, attention is focused almost entirely on it and on age distributions confined to the prereproductive and reproductive age groups.

The matrix A is of order $(k+1)\times(k+1)$, where $x = k$ is the last age group in which reproduction occurs, and written in full,

$$A = \begin{bmatrix} F_0 & F_1 & F_2 & F_3 & \dots & F_{k-1} & F_k \\ P_0 & . & . & . & \dots & . & . \\ . & P_1 & . & . & \dots & . & . \\ . & . & P_2 & . & \dots & . & . \\ \multicolumn{7}{c}{\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots} \\ . & . & . & . & \dots & P_{k-1} & . \end{bmatrix}.$$

This matrix is non-singular, since the determinant $|A| = (-1)^{k+2}(P_0P_1P_2\dots P_{k-1}F_k)$. There exists, therefore, a reciprocal matrix of the form

$$A^{-1} = \begin{bmatrix} . & P_0^{-1} & . & . & \dots & . \\ . & . & P_1^{-1} & . & \dots & . \\ . & . & . & P_2^{-1} & \dots & . \\ \multicolumn{6}{c}{\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots

5. Transformation of the co-ordinate system

Hitherto an age distribution n_{xt} has been regarded as a matrix consisting of a single column of elements. For simplicity in notation, this column vector will now be termed the vector ξ and different ξ's will be distinguished by different subscripts (ξ_a, ξ_x, etc.). We may picture an age distribution as a vector having a certain magnitude and related to a definite direction in a vector space, the space of the ξ's. The different age distributions which may arise in the case of any particular population will be assumed to be ξ's all radiating from a common origin. The numerical elements of a ξ vector are thus taken to be the co-ordinates of a point in multi-dimensional space referred to a general Cartesian co-ordinate system, in which the reference axes may make any angles with one another. At this point in the argument another type of vector will be introduced, which in matrix notation will be written as a row vector, and which will be termed the vector η. There is an intimate relationship between this new type and the old, for, associated with each vector ξ_a, there is a uniquely determined vector η_a, and vice versa. The inner or scalar product, $\eta_a \xi_a$, is the square of the length of the vector ξ_a. Either we may picture each of these vectors as associated with a different kind of vector space, the space of the ξ's and the dual space of the η's, which are not entirely disconnected but related in a special way; or, alternatively, we may regard them as two different kinds of vector associated with the same vector space. The relationship between η and ξ is precisely the same as that between covariant and contravariant vectors in differential geometry.

If we pass from our original co-ordinate system to a new frame of reference, and the variables η and ξ undergo the non-singular linear transformations,

$$\eta = \phi H, \quad \xi = H^{-1}\psi, \quad |H| \neq 0,$$

it can be seen that since the variables are contragredient, $\eta\xi = \phi\psi$, so that the square of the length of a vector remains invariant. Moreover, since the result of operating on a vector ξ_a with the matrix A is, in general, another vector ξ_b, where ξ_a and ξ_b are both referred to the original co-ordinate system, it follows that in the new frame of reference which is defined by the linear transformations given above, the relationship

$$A\xi_a = \xi_b$$

becomes

$$HAH^{-1}\psi_a = \psi_b,$$

or

$$B\psi_a = \psi_b.$$

Thus, in the new frame of reference the matrix $B = HAH^{-1}$ operating on the vector ψ_a is equivalent to the matrix A operating on the vector ξ_a in the original frame.

It is convenient, for the purposes of studying the matrix A and of performing any numerical computations with it, to transform the variables η and ξ in the above way, choosing the matrix H so as to make $B = HAH^{-1}$ as simple as possible. For $B^t = (HAH^{-1})^t = HA^tH^{-1}$, and since A is non-singular, by the reversal law, $(HAH^{-1})^{-1} = HA^{-1}H^{-1}$. Thus, if $f(A)$ is a rational integral function of A, $f(B) = f(HAH^{-1}) = Hf(A)H^{-1}$; and the properties of matrix functions $f(A)$ can be studied by means of the simpler forms $f(B)$. Moreover, the matrices A and B have the same characteristic equation and, therefore, the same latent roots. For $B - \lambda I = H(A - \lambda I)H^{-1}$ and, forming the determinants of both sides,

$$|B - \lambda I| = |H|\,|A - \lambda I|\,|H|^{-1},$$

so that the characteristic equation is

$$|A - \lambda I| = |B - \lambda I| = 0.$$

If, in the present case, the transforming matrix is taken to be

$$H = \begin{bmatrix} (P_0P_1P_2 \dots P_{k-1}) & . & . & \dots & . & . & . & . \\ . & (P_1P_2P_3 \dots P_{k-1}) & . & \dots & . & . & . & . \\ . & . & (P_2P_3 \dots P_{k-1}) & \dots & . & . & . & . \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ . & . & . & \dots & (P_{k-2}P_{k-1}) & . & . & . \\ . & . & . & \dots & . & P_{k-1} & . & . \\ . & . & . & \dots & . & . & 1 & . \end{bmatrix}$$

in which, it is to be noted, the only numerical elements lie in the principal diagonal and are derived entirely from the life table, then

$$B = HAH^{-1} = \begin{bmatrix} F_0 & P_0F_1 & P_0P_1F_2 & P_0P_1P_2F_3 & \dots & (P_0P_1P_2 \dots P_{k-1})F_k \\ 1 & . & . & . & \dots & . \\ . & 1 & . & . & \dots & . \\ . & . & 1 & . & \dots & . \\ . & . & . & 1 & \dots & . \\ \dots & \dots & \dots & \dots & \dots & \dots \\ . & . & . & . & \dots 1 & . \end{bmatrix}.$$

Comparing this matrix B with the original form A, it can be seen that the latter has been simplified to the extent that the original P_x figures in the principal subdiagonal are now replaced by a series of units, and the matrix A has been reduced to the rational canonical form $B = HAH^{-1}$ (see Turnbull & Aitken, 1932, chap. v). In this way any computations with the matrix A are made easier, and we may work henceforward in terms of ϕ and ψ vectors together with the matrix B, instead of with the original η and ξ vectors, and the matrix A. Any results obtained in this new system of co-ordinates may be transformed back again to the original system whenever necessary. It is evident that by suitably enlarging H the original matrix M may be transformed in a similar way.

This linear transformation of the original co-ordinate system is equivalent biologically to the transformation of the original population we were considering into a new and completely imaginary type which, although intimately connected with the old, has certain quite different properties. Thus, it can be seen from the transformed matrix B that the individuals in this new population, instead of dying off according to age as the original ones did, live until the whole span of life is completed, when they all die simultaneously. This is indicated by the P_x figures being now all equal to unity; an individual alive in the age group x to $x+1$ at $t = 0$ is certain of being alive at $t = 1$, excepting in the last age group of all where none of the individuals will be alive in one unit's time. Accompanying this somewhat radical change in the life table, there is a compensatory adjustment made in the rates of fertility so that the new population has the same inherent power of natural increase (r) as that of the old. This follows from the fact that the latent roots of the matrices A and B are the same, and, as will be shown later, the dominant latent root is closely related to the value of r obtained by the usual methods of computation. Insomuch as the transformation is reversible and $A = H^{-1}BH$, it can be seen that by changing H we could transform the canonical form B, if we wished, into another matrix in which the P_x subdiagonal might be a specified set of

figures derived from some other form of life table. But, for our present purposes, the canonical form B, in which all the P_x figures are units, offers advantages over any other matrix of a similar type owing to the greater ease with which it can be handled.

6. Relation between the canonical form B and the $L_x m_x$ column

The actual computation of the matrix B by way of the steps indicated in the theoretical development is by no means difficult, although it is a somewhat tedious process, particularly if the matrix is of a large order. The numerical elements in the first row of B for the brown rat are given in Table 5 of the Appendix. These values were obtained from the F_x and P_x figures which have already been used in the numerical example in § 3 and which will be found in the same table. Further reflection suggested, however, that instead of first of all obtaining A and then transforming to B, a short cut could be taken which would save labour and which also would tend to eliminate some of the small cumulative errors arising in the longer method.

The series of values P_0, P_0P_1, $P_0P_1P_2$, ..., $(P_0P_1P_2 \dots P_{k-1})$ by which the individual F_x figures are multiplied in order to obtain the first row of B, is essentially a stationary age distribution. For, since by definition,

$$P_x = \frac{L_{x+1}}{L_x},$$

$$(P_0P_1P_2 \dots P_x) = \frac{L_{x+1}}{L_0},$$

where $L_0 = \int_0^1 l_x dx$. Hence the required series of multipliers is given by a stationary age distribution in which only one individual is alive in the age group 0–1. Now, the F_x figures, as defined in § 2, already contain within them some allowance not only for the probability of survival during the first unit of life, but also for the fact that some adult individuals in each age group are dying off during the interval of time 0–1. The process of multiplying F_x by $(P_0P_1P_2 \dots P_{x-1})$ is thus analogous to the formation of the $L_x m_x$ column, by means of which the net reproduction rate is estimated. The chief difference between the first row of B and the $L_x m_x$ distribution is that in the former the maternal frequency is expressed as between the ages of $x+\frac{1}{2}$ to $x+1\frac{1}{2}$, instead of between x to $x+1$ as in the latter. If each element $(P_0P_1P_2 \dots P_{x-1}F_x)$ of the first row of B is regarded as centred at the age of $x+1$, the sum, mean and seminvariants of this 'distribution' may be estimated and compared with the values which are obtained from the $L_x m_x$ column in the process of calculating r by the usual methods. In the present numerical example the results of this comparison were as follows:

Parameter	$L_x m_x$ column	First row of B
Sum (R_0)	25·65786	25·6603
Mean	9·60604	9·5948
m_2	14·14397	14·1839
m_3	22·15696	21·9358
$m_4 - 3m_2^2$	−117·6480	−117·920

After allowing for the small cumulative errors which might be expected to occur in the calculation of the matrix elements, there is a substantial agreement between the respective

estimates. This agreement strongly suggests that if we had wished to pass immediately to the matrix B without going through the laborious process of calculating the F_x and P_x figures, the elements of the first row could have been obtained by forming a new $L_x m_x$ column in which the age group limits were shifted a half unit later in life. This could readily be done by interpolating in the integral curve of the $L_x m_x$ values for the ages $x+\frac{1}{2}$. This method of forming the first row of B has been adopted in other instances, when the matrix A was not of any immediate interest. It proved to be relatively quick and certainly less laborious than the method of first establishing A and then transforming to B which was the one used in the present numerical example.

7. THE STABLE AGE DISTRIBUTION

The result of operating on an age distributon ψ_x with the matrix B is, in general, a different distribution ψ_y. But, in the special case when the relation between the two distributions is such that

$$B\psi_a = \lambda\psi_a,$$

where λ is an algebraic number, then ψ_a may be said to be a stable age distribution appropriate to the matrix B. For the sake of brevity it will be referred to as a stable ψ. Similarly for initial row vectors, if

$$\phi_a B = \lambda\phi_a,$$

then ϕ_a is said to be a stable ϕ.

The matrix equation defining a stable ψ may be written as $k+1$ linear equations, of which the ith is

$$\sum_{j=1}^{k+1} b_{ij}n_j - \lambda n_i = 0,$$

where n_i $(i = 1, 2, \ldots, k+1)$ are the co-ordinates of the stable ψ, and b_{ij} the element in the ith row and jth column of B. Eliminating the n_i from this system of equations, we obtain the characteristic equation of B, namely,

$$|B - \lambda I| = 0;$$

and, expanding this determinant in powers of λ, we have in the present case,

$$\lambda^{k+1} - F_0\lambda^k - P_0F_1\lambda^{k-1} - P_0P_1F_2\lambda^{k-2} - \ldots - (P_0P_1\ldots P_{k-2})F_{k-1}\lambda - (P_0P_1\ldots P_{k-1})F_k = 0.$$

The $k+1$ roots λ_a of this equation are the latent roots of B, and corresponding to each distinct λ_a there is a pair of stable vectors, ϕ_a and ψ_a, determined except for an arbitrary scalar factor.

Once a latent root λ_a has been determined, it is a comparatively simple matter to find the appropriate stable ψ_a and ϕ_a vectors. Thus, it is easily shown that the stable ψ_a is the column vector $\{\lambda_a^k \lambda_a^{k-1} \lambda_a^{k-2} \ldots \lambda_a 1\}$. A short method of estimating ϕ_a is the following. Suppose, to take a simple case, that

$$B = \begin{bmatrix} a & b & c & d \\ 1 & . & . & . \\ . & 1 & . & . \\ . & . & 1 & . \end{bmatrix}$$

and let y_x $(x = 1, 2, 3, 4)$ be the elements of the stable ϕ_a appropriate to the root λ_a. Then

$$\begin{aligned} \phi_a B &= [ay_1 + y_2 \quad by_1 + y_3 \quad cy_1 + y_4 \quad dy_1] \\ &= [\lambda_a y_1 \quad \lambda_a y_2 \quad \lambda_a y_3 \quad \lambda_a y_4]. \end{aligned}$$

By equating similar elements and putting $y_1 = 1$, $y_4 = d/\lambda_a$, $y_3 = \dfrac{c+y_4}{\lambda_a}$, etc., it is easy to see how the required row vector can be built up. Having in this way obtained the stable ψ and ϕ vectors for the matrix B, they may be transformed to the appropriate stable ξ and η for the matrix A by means of the relations

$$\eta = \phi H, \quad \xi = H^{-1}\psi.$$

The characteristic equation of the matrix B, when expanded, is of degree $k+1$ in λ, and once B has been obtained this equation can immediately be written down, since the numerical coefficients of $\lambda^k, \lambda^{k-1}, \lambda^{k-2}$, etc., are merely the elements of the first row taken with a negative sign. Since there is only one change of sign in this equation, only one of the latent roots will be real and positive. Excluding the rather special case when the first row of B has only a single non-zero element, and taking the more usual type of matrix which will be met with, namely, that for a species breeding continuously over a large proportion of its total life span, it will be found that the modulus of this root (λ_1) is greater than any of the others,

$$|\lambda_1| > |\lambda_2| > |\lambda_3| > \ldots > |\lambda_{k+1}|,$$

the remaining roots being either negative or complex.

This dominant latent root λ_1, which will be $\gtreqless 1$ according as to whether the sum of the elements in the first row of B is $\gtreqless 1$, is the one which is principally of interest. Since it is real and positive, it is the only root which will give rise to a stable ψ or ξ vector consisting of real and positive elements. It is this stable ξ_1 associated with the dominant root λ_1 which is ordinarily referred to as the stable age distribution appropriate to the given age specific rates of fertility and mortality. Since

$$A^t\xi_1 = \lambda_1^t\xi_1,$$

it can be seen that the latent root λ_1 of the matrix A and the value of r obtained in the usual way from

$$\int_0^\infty e^{-rx} l_x m_x dx = 1,$$

are related by

$$\log_e \lambda_1 = r.$$

From the mathematical point of view, however, the negative and complex roots of the characteristic equation are of importance in the further theoretical development. Moreover, as will be shown later, the stable vectors associated with them are not entirely without interest. Two main cases then arise: when the remaining roots are all distinct, and when there are repeated roots. For the present it will be assumed that the latent roots of the matrix are all distinct.

8. Properties of the stable vectors

Before proceeding further it is necessary to mention briefly the reasons why the methods given above for the computation of the stable ψ and ϕ vectors were adopted, apart from their simplicity in practice. If the $k+1$ distinct roots of the characteristic equation are known, we may form a set of $k+1$ matrices $f(\lambda_a)$ by inserting in turn the numerical value of each root in the matrix $[B-\lambda_a I]$. The adjoint of $f(\lambda_a)$ is

$$F(\lambda_a) = \prod_{b \neq a} [B-\lambda_b I] \quad \text{and} \quad f(\lambda_a) F(\lambda_a) = 0.$$

It may be shown that the stable ψ_a appropriate to the root λ_a can be taken proportional to any column, and the stable ϕ_a proportional to any row of the matrix $F(\lambda_a)$ (see e.g. Frazer, Duncan & Collar, 1938, chap. III). Moreover, $F(\lambda_a)$ is a matrix product of the type $\psi\phi$,

where the ψ vector is given by the first column and the ϕ vector by the last row of $F(\lambda_a)$, each divided by the square root of the element in the bottom left-hand corner; and the trace of the matrix is equal to the scalar product $\phi\psi$. Now $[B-\lambda_a I]$ is a square matrix of order $k+1$ with only zero elements below and to the left of the principal subdiagonal, which itself consists of units. The product of k such matrices, which gives $F(\lambda_a)$, will have therefore a unit in the bottom left-hand corner. Since the stable ϕ_a and ψ_a vectors obtained by the methods suggested in § 7 have respectively their first and last elements $= 1$, it follows that

$$\psi_a\phi_a = F(\lambda_a), \quad \phi_a\psi_a = \text{trace } F(\lambda_a).$$

The stable vectors may now be normalized. If the scalar product, $\phi_a\psi_a = z^2$, say, then

$$\frac{\phi_a}{|z|}\frac{\psi_a}{|z|} = 1.$$

From now on it will be assumed that the stable vectors appropriate to each of the latent roots have been normalized in this way.

These vectors have the following important properties:

(1) The $k+1$ stable ψ are linearly independent. There is thus no such relationship, with non-zero coefficients c, as

$$c_1\psi_1+c_2\psi_2+c_3\psi_3+\ldots+c_{k+1}\psi_{k+1} = 0.$$

(2) The scalar product of a stable ψ, ψ_a with the associated vector of another stable ψ, ψ_b is zero, i.e.

$$\phi_b\psi_a = 0 \quad (a \neq b).$$

The normalized stable ψ thus form a set of $k+1$ independent and mutually orthogonal vectors of unit length.

(3) Any arbitrary ψ—ψ_x say—can be expanded in terms of the stable ψ, thus

$$\psi_x = c_1\psi_1+c_2\psi_2+c_3\psi_3+\ldots+c_{k+1}\psi_{k+1},$$

where the coefficients c may be either real or complex. Similarly an arbitrary vector ϕ_x can be expanded in terms of the stable ϕ.

9. The spectral set of operators

The matrix product $\psi_a\phi_a$ of the normalized stable vectors associated with the latent root λ_a will be termed the matrix S_a. From the relationships which have already been given, it can be seen that S_a is merely the adjoint matrix $F(\lambda_a)$ of the previous section after each element in the latter has been divided by the sum of the elements in the principal diagonal; in other words it is the normalized $F(\lambda_a)$. In the case of all the latent roots being distinct, there are thus $k+1$ matrices S_a, and these S_a form a spectral set of operators with the following properties:

$$S_a^2 = S_a, \quad S_aS_b = 0 \quad (a \neq b), \quad \sum_{a=1}^{k+1} S_a = I.$$

Moreover, if $f(B)$ is a polynomial of the matrix B, we have by Sylvester's theorem (Turnbull & Aitken, 1932, chap. VI, § 8)

$$f(B) = \sum_{a=1}^{k+1} f(\lambda_a)\, S_a,$$

so that the matrix

$$B = \lambda_1S_1+\lambda_2S_2+\ldots+\lambda_{k+1}S_{k+1}, \quad \text{and} \quad B^t = \lambda_1^tS_1+\lambda_2^tS_2+\ldots+\lambda_{k+1}S_{k+1}.$$

If the latent roots in the expansion of B are raised to a high power, the term associated with the positive real root predominates over all the others, so that when t is large, we have approximately

$$B^t = \lambda_1^t S_1.$$

In any particular case the power to which B will have to be raised in order that this equation should be approximately true, will depend both on the order of the matrix and on the relative magnitude of the dominant root as compared with that of the remaining roots.

At this point it is possible to attach some biological meaning to one of the ϕ or η row vectors, which in the first place were introduced into the theory for reasons of symmetry, and which were defined solely in terms of their mathematical properties. If at a given moment a transformed population has an arbitrary age distribution ψ_x, and the sequence $B\psi_x, B^2\psi_x, \ldots, B^t\psi_x$ is formed, it can be seen that when t is large and ψ_x is expanded in terms of the stable ψ, we have approximately

$$B^t\psi_x = c_1\lambda_1^t\psi_1.$$

Thus, a population with any arbitrary age distribution tends ultimately to approach the stable form appropriate to the given rates of fertility and mortality, provided that these age-specific rates remain constant. This theorem is, of course, well known; and it is clear that the achievement of the stable form of age distribution associated with the dominant latent root is very unlikely to occur in practice, except in the case when the initial distribution is already of that form or exhibits only small departures from it. Now, it has already been shown that the sums of the columns of a matrix B^t provide a measure of the contributions made to the population at time t per individual alive in the respective age groups at $t = 0$. When t is large, the matrix B^t is equivalent to the matrix S_1 multiplied by a scalar factor. From the way in which this latter matrix was constructed by the outer multiplication of ψ_1 and ϕ_1, it is evident that the sums of the columns of S_1 are proportional to the vector ϕ_1. Thus, transforming back again to the original co-ordinate system, the stable η_1 associated with the dominant latent root provides a measure of the relative contributions per head made to the stable population by the individual age groups.

10. Reduction of B to classical canonical form

From the $k+1$ stable ψ a matrix Q can be constructed, whose columns are the stable ψ arranged, reading from left to right, in descending order of the moduli of the roots with which they are associated. Corresponding to every pair of complex roots, $u \pm iv$, there will be in this matrix a pair of columns consisting of complex elements, the one column being the conjugate complex of the other. Some of the columns associated with the negative roots may be purely imaginary owing to the normalization of the corresponding ψ and ϕ vectors. In a similar way a matrix U may be formed, whose rows, reading from above down, are the stable ϕ arranged in the same order. Since the stable ϕ and ψ are normalized, and $\phi_a\psi_b = 0$ for $a \neq b$,

$$UQ = I,$$

and, therefore, U and Q are reciprocal matrices. By premultiplying and postmultiplying respectively with U and Q, the matrix B may be reduced to the classical canonical form C, in which the only elements lie in the principal diagonal and consist of the latent roots arranged in the order prescribed above. This reduction of B to a purely diagonal form by means of the collineatory transformation $UBQ = C$ is, however, only possible in the type of matrix we are considering, when the latent roots are all distinct.

The expansion of an arbitrary ψ_x in terms of the stable ψ,

$$\psi_x = c_1\psi_1 + c_2\psi_2 + \ldots + c_{k+1}\psi_{k+1},$$

may be written in matrix notation as $\psi = Qc$,

where c is the column vector $\{c_1 c_2 c_3 \ldots c_{k+1}\}$. Similarly, the expansion of the vector ϕ_x associated with ψ_x may be written $\phi = dU$,

where d is the row vector $[d_1 d_2 d_3 \ldots d_{k+1}]$. This is again a transformation to another co-ordinate system, but this time the reference axes are at right angles to one another. Since the variables transform contragrediently, $dc = \phi\psi = \eta\xi$. At this point it is necessary to make some assumption as to the relationship between the elements of the vectors d and c. Since these elements may be either real or complex, it will be assumed that

$$d = \bar{c}',$$

where the row vector $\bar{c}'$ is the transposed conjugate complex of the column vector c. Hence, the square of the length of a vector referred to this orthogonal co-ordinate system is given by $\bar{c}'c$, a number which is essentially real and non-negative. (The assumption that $d = c'$ will be found, in the particular case studied here, to lead to values of $c'c$ which, although real, may be negative.)

11. THE RELATION BETWEEN ϕ AND ψ VECTORS

Since $\psi_x = Qc_x$, and the associated $\phi_x = \bar{c}'_x U$, it may be seen from the relations given in the two previous sections that $\bar{\phi}'_x = \bar{U}'U\psi_x = G\psi_x$.

The matrix $G = \bar{U}'U$ is symmetrical and all its elements are real numbers, those in the principal diagonal being necessarily positive in sign. It therefore remains unaltered after transposition. Since the elements of the vector ψ_x, which is by definition an age distribution transformed by the matrix H, are also necessarily real, we may write

$$\phi_x = \psi'_x G.$$

The role of the matrix G is therefore the same as that of the double covariant metric tensor g_{mn} in the tensor calculus. It transforms any ψ vector into its associated ϕ vector. This process is reversible, the reciprocal matrix being given by $G^{-1} = Q\bar{Q}'$.

The magnitude of a vector ψ_x is defined by the equation

$$x = (\psi'_x G\psi_x)^{\frac{1}{2}},$$

where the square root is taken with a positive sign. If we have two vectors ψ_x and ψ_y, both radiating from the common origin, the angle between them is given by

$$\cos\theta = \frac{\psi'_x G\psi_y}{xy},$$

from which it follows that when $\psi'_x G\psi_y = \phi_x\psi_y = 0$, the two vectors are at right angles to one another, and when $\cos\theta = 1$ their directions are the same. If, in the last equation, we take ψ_y to be the stable vector ψ_1 associated with the dominant latent root λ_1, then knowing the magnitude of a vector $B^t\psi_x$ and the angle which it makes with the ψ_1 axis, we can obtain a graphical representation of the way in which a particular age distribution approaches the stable form.

The matrix G also defines the angles between the reference axes of the co-ordinate system. If we introduce into the vector space of the ψ's a system of reference axes defined by the unit column vectors

$$\begin{aligned} e_1 &= \{1 \quad 0 \quad 0 \quad \dots \quad 0\} \\ e_2 &= \{0 \quad 1 \quad 0 \quad \dots \quad 0\} \\ e_3 &= \{0 \quad 0 \quad 1 \quad \dots \quad 0\} \\ &\vdots \\ e_{k+1} &= \{0 \quad 0 \quad 0 \quad \dots \quad 1\} \end{aligned}$$

then the distance from the origin of the unit point e_j is

$$Oe_j = \sqrt{g_{jj}},$$

and the angle between any two of the co-ordinate axes is given by

$$\cos\theta_{ij} = \frac{g_{ij}}{\sqrt{(g_{ii}g_{jj})}},$$

where, in both cases, g_{ij} is the element in the ith row and jth column of G.

By transforming back to the original co-ordinate system, the metric matrix associated with the vector space of the ξ's will be found to be

$$G_\xi = HG_\psi H.$$

Hitherto we have been chiefly concerned with an operator B which, acting in the vector space of the ψ's, has the power of transforming a vector ψ_x into what is in general a new vector ψ_y. We may now have reason to inquire how the associated vectors ϕ_x and ϕ_y are related in the vector space of the ϕ's, whenever

$$B\psi_x = \psi_y.$$

Since $$\psi_x = G^{-1}\phi'_x \quad \text{and} \quad \psi_y = G^{-1}\phi'_y,$$

we have $$GBG^{-1}\phi'_x = \phi'_y,$$

and hence, by transposing,

$$\phi_x G^{-1}B'G = \phi_y.$$

Thus, the matrix which transforms ϕ_x into ϕ_y is not the same as that which transforms ψ_x into ψ_y. In order to distinguish these two operators, they will be referred to as B_ϕ and B_ψ respectively. In the few numerical examples which have been worked out, the matrix B_ϕ differed greatly from the rational canonical form B_ψ, and consisted of $(k+1)^2$ real elements, some of which were negative. In addition to the relationship

$$B_\phi = G^{-1}B'_\psi G,$$

it was also found that in the case of distinct latent roots

$$B_\phi = \overline{\lambda}_1 S_1 + \overline{\lambda}_2 S_2 + \overline{\lambda}_3 S_3 + \dots + \overline{\lambda}_{k+1} S_{k+1},$$

where the S_a matrices are the spectral set of operators defined in § 9 and $\overline{\lambda}_a$ is the conjugate complex of the latent root λ_a. It may be seen from this expansion of B_ϕ that the necessary condition for $B_\phi = B_\psi$ is that all the latent roots of B should be real, the one positive and the remainder negative. (It is to be noted that we are dealing here with the case of distinct latent roots; it would appear that even if all the λ were real, $B_\phi \neq B_\psi$ in the case of repeated roots.) Unless, however, the matrix B is of a small order, it is unlikely that this condition would be fulfilled, since in the more usual-sized matrix we shall be dealing with in the case

of human or other mammalian populations, some of the roots will almost certainly be complex.*

It seems unlikely that the equations given in this section will have very much practical application at the moment; they have been included merely to fill in the picture of the relationship between the two types of vector. For all ordinary purposes no one would choose to work in terms of ϕ vectors and the operator B_ϕ, instead of the more obvious ψ vectors and the more simple matrix form B_ψ. Nevertheless, since it has been necessary to assume that there are such vectors as η or ϕ associated with every ξ or ψ, and since these vectors play such an important part in the mathematical theory, the question naturally arises as to what significance must be attached to them from the biological point of view. Have they in fact any real meaning at all? Or must they be regarded purely as mathematical abstractions? At the end of § 9 it has been suggested that the row vector associated with the stable ξ_1 appropriate to the dominant latent root is a measure of the contributions made to the stable population per individual female alive in the respective age groups of the initial distribution; but this is a special case and the interpretation offered here is not applicable, even in a wider form, to η vectors in general. It may well be, of course, that the latter as a class have no concrete meaning: and that in seeking to define them in terms of some property or characteristic of an age distribution one is merely attempting the impossible. But the fact of one η vector having been defined in non-mathematical terms, even though on further consideration some revision may be needed of the actual definition given here, suggests that impossible may perhaps be too final a word to use in this connexion.

12. Case of repeated latent roots

When any of the latent roots other than the real positive dominant root are repeated, a number of the relations given in the previous sections no longer hold good and certain equations must therefore be modified. Suppose a root λ_a has a multiplicity s, and consider the matrix $f(\lambda_a)$ such as, to take a simple example,

$$f(\lambda_a) = \begin{bmatrix} a-\lambda_a & b & c & d \\ 1 & -\lambda_a & 0 & 0 \\ 0 & 1 & -\lambda_a & 0 \\ 0 & 0 & 1 & -\lambda_a \end{bmatrix}.$$

Then, since the determinant $|f(\lambda_a)| = 0$ and at least one of the first minors of order 3 is not equal to zero, the above matrix has rank 3 and, therefore, nullity 1. Hence it can be seen that $f(\lambda_a)$ of whatever order it may be has nullity 1. Since $f(\lambda_a)$ is thus simply degenerate, there is only one stable ψ appropriate to the s equal roots λ_a (see e.g. Frazer, Duncan & Collar, 1938, chap. III).

Certain consequences immediately follow. Since the matrices Q and U cannot be constructed in the way given in § 10, the reduction of B to a purely diagonal matrix by means of the collineatory transformation UBQ can no longer be carried out. Neither is the expansion of B in terms of the spectral set of S_a matrices, nor the expansion of an arbitrary ψ_x

* The interesting theoretical case of the matrix A or B having a number of its latent roots real and positive, with the remainder real and negative, is outside the scope of the present study. The necessary conditions for this to be true would involve a number of the F_x figures becoming negative, a case not considered here, but which biologically might be held to correspond with the destruction of eggs, or the very young, by certain age groups, e.g. as observed by Chapman (1933) in experimental populations of the flour beetle, *Tribolium confusum*.

in terms of the stable ψ, possible in the actual forms given in §§ 9 and 8. We may, however, proceed as follows.

The matrix Q is essentially an alternant which has been postmultiplied by a diagonal matrix N, the elements in the latter being given by the reciprocals of the scalar factors $|z|$ by which the stable vectors were divided in the process of normalization $\left(\frac{\phi_a}{|z|}\frac{\psi_a}{|z|} = 1\right)$, so that $Q = XN$, where

$$X = \begin{bmatrix} \lambda_1^k & \lambda_2^k & \dots & \lambda_{k+1}^k \\ \lambda_1^{k-1} & \lambda_2^{k-1} & \dots & \lambda_{k+1}^{k-1} \\ \vdots & \vdots & & \vdots \\ \lambda_1 & \lambda_2 & \dots & \lambda_{k+1} \\ 1 & 1 & \dots & 1 \end{bmatrix}.$$

When a root λ_a is repeated s times, s of the columns in X become the same, and therefore the matrix becomes singular. In place of this alternant matrix we have the confluent alternant form (see Turnbull & Aitken, 1932, chap. VI) in which the s columns ($s = 0, 1, 2, 3, \dots, s-1$) corresponding to the repeated root λ_a are written

$$\begin{array}{llll} \lambda_a^k & k\lambda_a^{k-1} & \frac{k(k-1)\lambda_a^{k-2}}{2!} & \dots \\ \lambda_a^{k-1} & (k-1)\lambda_a^{k-2} & \frac{(k-1)(k-2)}{2!}\lambda_a^{k-3} & \dots \\ \vdots & \vdots & \vdots & \dots \\ \vdots & \vdots & 3\lambda_a & \dots \\ \vdots & 2\lambda_a & 1 & \dots \\ \lambda_a & 1 & 0 & \dots \\ 1 & 0 & 0 & \dots \end{array}$$

the column s being obtained from column 0 (the non-normalized stable ψ_a) by the operation $\left(\frac{d}{d\lambda_a}\right)^s \Big/ s!$. This confluent alternant form of X is non-singular and therefore a reciprocal matrix can be determined (see Aitken, 1939, § 50). The general classical canonical form of B obtained by the collineatory transformation, $X^{-1}BX = C$, has corresponding to the repeated root λ_a a diagonal submatrix:

$$\begin{bmatrix} \lambda_a & 1 & . & . & . & . \\ . & \lambda_a & 1 & . & . & . \\ . & . & \lambda_a & 1 & . & . \\ \dots & \dots & \dots & \dots & \dots & \dots \\ . & . & . & . & \lambda_a & 1 \\ . & . & . & . & . & \lambda_a \end{bmatrix}.$$

The matrix product of a column of X with the appropriate row of the reciprocal X^{-1} gives as before S_a and the $k+1$ S_a form a spectral set with the same properties as those defined in § 9, except that

$$B \neq \sum_{a=1}^{k+1} \lambda_a S_a.$$

In place of this expansion of B in terms of the S_a matrices, we have in the case of repeated roots the confluent form of Sylvester's theorem, for details of which reference may be made to Frazer, Duncan & Collar, 1938, chap. III. Apart from this modification, however, we may obtain by inspection of the S_a the factors by which the respective columns of X must be divided in order to express this matrix in a form comparable to that of Q in § 10. Similarly, when the respective rows of X^{-1} are multiplied by the appropriate factors, the matrix U is found and hence $G = \overline{U}'U$ can be constructed. (It is to be noted that $(\overline{X^{-1}})' X^{-1}$ is neither equal to, nor directly proportional to $\overline{U}'U$.) An arbitrary ψ_x can be expanded in terms of the column vectors of Q, though in the case of only one column associated with the repeated root λ_a does the relationship $B\psi_a = \lambda_a \psi_a$ hold.

13. THE APPROACH TO THE STABLE AGE DISTRIBUTION

A stable age distribution appropriate to the matrix B has been defined mathematically by the equation

$$B\psi = \lambda\psi,$$

and it has already been shown that since only one latent root of B is real and positive, only one of the stable ψ will consist of real and positive elements. But, in addition to this Malthusian age distribution, it is also of some interest to inquire whether any significance can be attached to the remaining stable ψ associated with the negative and complex roots of the characteristic equation.

Any age distribution ψ_x, the elements of which are necessarily $\geqq 0$, may be expressed as a vector of deviates from the stable ψ_1 associated with the dominant latent root, and we may therefore write the expansion of ψ_x in terms of the stable ψ as

$$(\psi_x - c_1\psi_1) = c_2\psi_2 + c_3\psi_3 + \ldots + c_{k+1}\psi_{k+1} = \psi_d,$$

where the coefficients c are given by the vector $c = U\psi_x$. Thus, the way in which a particular type of age distribution will approach the stable form may be studied by means of the vector ψ_d.

Among the terms occurring in the right-hand side of this expression there will be, corresponding to each negative root, a single term $c_a\psi_a$ which will consist of real elements alternately positive and negative in sign. (Even if the normalized ψ_a is imaginary this term will consist of real numbers, since in this case c_a will also become imaginary.) Moreover, corresponding to every pair of complex roots there will be a pair of terms $(c_m\psi_m + c_n\psi_n)$ which taken together will also give a single vector with real elements. This follows from the fact that c_m is the conjugate complex of c_n owing to the way in which the matrix U is constructed. Then, apart from the scalar c_1 which must necessarily be > 0, some of the remaining coefficients $c_2, c_3, \ldots, c_{k+1}$ in the expansion of ψ_d may be zero. The first and most obvious case is when they are all zero, and the age distribution ψ_x is therefore already of the stable form. But, if either

$$\psi_d = c_a\psi_a,$$

where ψ_a corresponds to a negative latent root, or

$$\psi_d = c_m\psi_m + c_n\psi_n,$$

where ψ_m and ψ_n are associated with a conjugate pair of complex roots, then it follows that the age distribution ψ_x will, as time goes on, approach the stable form in a particular way defined by either

$$B^t\psi_d = c_a\lambda^t\psi_a \quad \text{or} \quad B^t\psi_d = c_m\lambda^t\psi_m + c_n\overline{\lambda}^t\psi_n,$$

in which λ^t for a pair of complex roots $u+iv$ with modulus r may be written in the form of $r^t(\cos\theta t \pm i \sin\theta t)$. Thus, the negative and complex latent roots of B serve to determine a number of age distributions which are of some interest owing to the fact that they will approach the Malthusian form in what may be termed a stable fashion.

Since $|\lambda_1| > |\lambda_2| > |\lambda_3| > \ldots > |\lambda_{k+1}|$, the vector of deviates ψ_d will tend towards zero as $t \to \infty$ whenever $\lambda_1 \gtrless 1$. Thus, in the case of a stationary population, any ψ_x will converge to the stable form of age distribution. But if $\lambda_1 > 1$, there is a possibility of one or more of the remaining roots having a modulus $\geqq 1$, e.g. $|\lambda_2| \geqq 1$. In the latter case there may be certain age distributions with $c_2 \neq 0$ for which the amplitude of the deviations from the stable form tend either to increase ($|\lambda_2| > 1$), or to remain constant ($|\lambda_2| = 1$). From the practical point of view, however, we may still say that a population with such an age distribution approaches or becomes approximately equal to the stable population, since λ_1^t is much greater than λ_2^t when t is large.

14. Special case of the matrix with only a single non-zero F_x element

The interesting case of the matrix A having only a single non-zero element in the first row has been illustrated in a numerical example by Bernardelli (1941).* This author has also used a matrix notation in the mathematical appendix to his paper, and the form of his basic matrix is the same as that referred to here as M or A. It is not clear, however, from the definitions which he gives whether he regards the elements in the first row of his matrix as being constituted by the maternal frequency figures (m_x) themselves, or by a series of values similar to those defined here as the F_x figures. He refers to them merely as the specific fertility rates for female births.

In discussing the causes of population waves, Bernardelli describes a hypothetical species, such as a beetle, which lives for only three years and which propagates in the third year of life. He assumes, for the sake of argument, that—to employ the terminology used here—$P_0 = \frac{1}{2}$ and $P_1 = \frac{1}{3}$, and that 'each female in the age 2–3 produces, on the average, 6 new living females'. Assuming for the moment that he is here defining a F_x figure, we may write this system of mortality and fertility rates as

$$A = \begin{bmatrix} 0 & 0 & 6 \\ \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{3} & 0 \end{bmatrix}, \quad B = HAH^{-1} = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}.$$

The characteristic equation expanded in terms of λ is $\lambda^3 - 1 = 0$; and the latent roots are therefore 1, $-\frac{1}{2} \pm \frac{\sqrt{3}}{2} i$, all three being of equal modulus. The matrix A has the interesting properties

$$A^2 = A^{-1}, \quad A^3 = I,$$

so that any initial age distribution repeats itself regularly every three years. Thus, as Bernardelli shows, a population of 3000 females distributed equally among the three age

* At this point I should like to acknowledge the gift of a reprint of this paper, which was received by the Bureau of Animal Population at a time when I was in the middle of this work, and when I was just beginning to appreciate the interesting results which could be obtained from the use of matrices and vectors: also a personal communication from Dr Bernardelli, received early in 1942, at a time when it was difficult to reply owing to the developments of the war situation in Burma. Although the problems we were immediately interested in differed somewhat, this paper did much to stimulate the theoretical development given here, and it is with great pleasure that I acknowledge the debt which I owe to him.

groups becomes a total population of 6833 at $t = 1$; of 5166 at $t = 2$; and again 3000 distributed equally among the age groups at $t = 3$. Unless a population has already an initial age distribution in the ratio of $\{6:3:1\}$, no approach will be made to the stable form associated with the real latent root, and the vector of deviates ξ_d will continue to oscillate with a stable amplitude, which will in part depend on the form of the initial distribution. Although this numerical example refers specifically to a stationary population, it is evident that a similar type of argument may be developed in the case when $|\lambda| > 1$ and $A^3 = \lambda^3 I$.

We have assumed here that his definition of the fertility rate refers to a F_x figure. But, if we were to interpret the words quoted above as referring to a maternal frequency figure, namely that every female alive between the ages 2–3 produces on the average 6 daughters per annum, then the results become entirely different. For, deriving the appropriate F_x figures by the method described in § 2, the matrix is now

$$A = \begin{bmatrix} 0 & 1 & 3 \\ \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{3} & 0 \end{bmatrix}, \quad B = HAH^{-1} = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix},$$

and the latent roots are $1, -\frac{1}{2} \pm \frac{1}{2}i$. The modulus of the pair of complex roots is $1/\sqrt{2}$, which is < 1, so that every age distribution will now converge to the stable form associated with the real root. Thus, to take the same example as before, 3000 females distributed equally among the age groups will tend towards a total population of 4000 distributed in the ratio of $\{6:3:1\}$, and it was found that this age distribution would be achieved at approximately $t = 23$. During the approach to this stable form periodic waves are apparent both in the age distribution and in the total number of individuals, but these oscillations are now damped, in contrast with the results obtained with the first type of matrix.

This simple illustration serves to emphasize the importance which must be attached to the way in which the basic data are defined and to the marked difference which exists between what are termed here the m_x and F_x figures. Nevertheless, apart from the question of the precise way in which the definition of the fertility rates is to be interpreted in this example, the first type of matrix with only a single element in the first row does correspond to the reproductive biology of certain species. Thus, in the case of many insect types the individuals pass the major portion of their life span in various immature phases and end their lives in a short and highly concentrated spell of breeding. The properties of this matrix suggest that any stability of age structure will be exceptional in a population of this type, and that even if the matrix remains constant we should expect quite violent oscillations to occur in the total number of individuals.

15. Numerical comparison with the usual methods of computation

From the practical point of view it will not always be necessary to estimate the actual values of all the stable vectors and of the associated matrices which are based on them. Naturally, much will depend on the type of information which is required in any particular case. In order to compute, for instance, the matrices U, Q and G, it is necessary first of all to determine all the latent roots of the basic matrix. The ease with which these may be found depends very greatly upon the order of the matrix. Thus, in the numerical example for the brown rat used previously in § 3, the unit of age and time is one month and the resulting square matrix A is of order 21. To determine all the 21 roots of the characteristic equation would be

a formidable undertaking. It might be sufficient in this case to estimate the positive real root and the stable vector associated with it. On the other hand, it is possible to reduce the size of the matrix by taking a larger unit of age, and in some types of problem, where extreme accuracy is not essential, a unit say three times as great might be equally satisfactory, which would reduce the matrix for the rat population to the order of 7×7. It is not too difficult to find all the roots of a seventh degree equation by means of the root-squaring method (Whittaker & Robinson, 1932, p. 106). But the reduction of the matrix in this way will generally lead to a value of the positive real root which is not the same as that obtained from the larger matrix, and it is therefore necessary to see by how much these values may differ owing to the adoption of a larger unit of time.

Another important point which must be considered is the following. By expressing the age specific fertility and mortality rates in the form of a matrix and regarding an age distribution as a vector, an element of discontinuity is introduced into what is ordinarily taken to be a continuous system. Instead of the differential and integral calculus, matrix algebra

Table 2

Age group (units of 30 days)	'True' stable age distribution	Matrix stable age distribution	Age group (units of 30 days)	'True' stable age distribution	Matrix stable age distribution
0–	37,440	37,362	12–	161	160
1–	22,595	22,644	13–	101	101
2–	14,417	14,444	14–	64	63
3–	9,227	9,238	15–	40	40
4–	5,903	5,906	16–	25	24
5–	3,775	3,775	17–	15	15
6–	2,413	2,412	18–	9	9
7–	1,542	1,540	19–	6	5
8–	984	982	20–	3	3
9–	627	626			
10–	399	398			
11–	254	253	Total	100,000	100,000

is used, a step which leads to a great economy in the use of symbols and consequently to equations which are more easily handled. Moreover, many quite complicated arithmetical problems can be solved with great ease by manipulating the matrix which represents the given system of age specific rates. But the question then arises whether these advantages may not be offset by a greater degree of inaccuracy in the results as compared with those obtained from the previous methods of computation. It is not easy, however, to settle this point satisfactorily. In the way the usual equations of population mathematics are solved, a similar element of discontinuity is introduced by the use of age grouping. Thus, in the case of a human population, if we were estimating the inherent rate of increase in the ordinary way, we should not expect to obtain the same value of r from the data grouped in five year intervals of age as that from the data grouped in one year intervals. The estimates of the seminvariants would not be precisely the same in both cases. Nevertheless, the estimate from the data grouped in five year intervals is usually considered to be sufficiently accurate for all ordinary purposes, and there is little doubt that if we merely require the inherent rate of increase and the stable age distribution, these methods of computation are perfectly satisfactory when applied to human data. But, in the case of rodents, and probably also

other species with high gross and net reproduction rates, it will be found that even a 4th degree equation in r with the coefficients based on the seminvariants of the $L_x m_x$ distribution is, in many examples, not sufficient to give an accurate estimate of the rate of increase, and it is necessary to arrive at a better value of r in a somewhat roundabout way. Here, the determination of the positive real root of the characteristic equation for the matrix, once the latter has been established, may be even quicker than finding a solution from the $L_x m_x$ column by a method such as that described in the appendix.

In order to compare the values of r obtained from the characteristic equation of the matrix with those obtained from the $L_x m_x$ column, both methods were used in the numerical example for the brown rat, and a comparison was also made between the values when the data were grouped in 1 month and in 3 month age intervals. In addition, the stable age distribution appropriate to the positive real root of the matrix was also calculated in both cases. The results were as follows.

(*a*) *One month unit of grouping; matrix of order* 21×21. Using the method of computation indicated in the appendix, the value of r was estimated to be 0·44565 per month of 30 days. The positive real root of the characteristic equation was $\lambda_1 = 1{\cdot}56246$, whence $r = 0{\cdot}44626$, a value which differs from the former only in the fourth decimal place. The appropriate age distributions, expressed per 100,000, are given in Table 2, the 'true' stable being obtained from

$$n_x = 100{,}000b\int_x^{x+1} e^{-rx} l_x \, dx.$$

The agreement between these distributions is very good, although the one derived from the matrix shows certain small rhythmical departures from the 'true' distribution particularly in the earlier age groups. The maximum difference between them in this region, however, is not greater than 2·2 per thousand. Since the matrix stable distribution is proportional to the columns of the matrix $H^{-1}S_1H$, which in turn is proportional to A^t when t is large, this agreement between the two distributions also indicates that the cumulative errors which might be expected in forming the series A^2, A^3, A^4, etc., owing to the P_x figures being based on the life table age distribution, are not very serious. Judging by this example it seems that satisfactory estimates of the inherent rate of increase and of the stable age distribution may be made from a large order matrix.

(*b*) *Three months unit of grouping; matrix of order* 7×7. Clearly there are several ways in which a large order matrix may be condensed into one of a smaller order. The method which was used in the present instance was to construct the first row of the condensed canonical form B by interpolating in the integral curve of the original $L_x m_x$ column (1 month units of grouping) for the ages 4·5, 7·5, 10·5, etc., and taking the first differences of the seven values thus obtained. Since interpolation was not very satisfactory in the earlier part of the integral curve—the differences converged rather slowly in this region—the elements were expressed to only three places of decimals. (Some preliminary transformation of the integral $L_x m_x$ figures might have been better in this case.) The characteristic equation, expanded in terms of λ, was found to be

$$\lambda^7 - 1{\cdot}756\lambda^6 - 6{\cdot}899\lambda^5 - 7{\cdot}203\lambda^4 - 5{\cdot}344\lambda^3 - 3{\cdot}244\lambda^2 - 1{\cdot}110\lambda - 0{\cdot}102 = 0.$$

It will be seen that the sum of the coefficients, $R_0 = 25{\cdot}658$, which is necessarily the same as the original net reproduction rate owing to the way in which the coefficients were derived. For interest, the seven roots of this equation were then determined by the root-squaring

method, using 4-place tables of logarithms and Barlow's tables of squares (Whittaker & Robinson, 1932, p. 110). The approximate values of the roots, arranged in descending order of their moduli, are

$$\lambda_1 = \ \ 4{\cdot}016,$$
$$\lambda_2 = -1{\cdot}032,$$
$$\lambda_3\lambda_4 = -0{\cdot}0215 \pm 0{\cdot}6762i \quad (\text{mod.} = 0{\cdot}6765),$$
$$\lambda_5\lambda_6 = -0{\cdot}5245 \pm 0{\cdot}3486i \quad (\text{mod.} = 0{\cdot}6298),$$
$$\lambda_7 = -0{\cdot}135.$$

Thus, apart from the positive real root, there are two negative and two pairs of complex roots. It is interesting to note in passing that in this example the modulus of the second latent root is >1 (vide § 13). From the value of the dominant root we find $r = 0{\cdot}4634$ per head per month of 30 days, an estimate of the rate of increase which is 1·71 % per month higher than that from the large order matrix.

Table 3. *Stable age distributions*

Age group (in months)	$100{,}000b\int_x^{x+3} e^{-rx} l_x dx$	From condensed matrix	Summation of 'true' distribution
0–	75,960	75,762	74,452
3–	18,055	18,267	18,905
6–	4,514	4,519	4,939
9–	1,120	1,111	1,280
12–	273	267	326
15–	64	61	80
18–	14	13	18
Total	100,000	100,000	100,000

The net reproduction rate given by the new $L_x m_x$ column which was obtained by working in units of three months, was 25·6162, a figure somewhat lower than the original one of 25·6579. The rate of increase, estimated in a similar way to the former example for one month age units, was $r = 0{\cdot}46034$, again a higher figure, though one of much the same order as that obtained from the condensed matrix.

The appropriate age distributions are given in Table 3, together with the 'true' distribution of Table 2 summed in three month age groups.

Compared with the last column, both of the stable distributions for the data grouped in three month age intervals are tilted towards the younger age classes, so that the number of immature females (<3 months of age) is overestimated, while the remaining age groups are underestimated. The distributions derived from the integral and from the matrix are again of much the same order, and the differences between them and the last column, although not very great, are quite marked.

The four estimates of the inherent rate of increase which have obtained from these numerical data may be compared in the following table.

In both cases the estimates from the $L_x m_x$ column and from the matrix agree very well: for a given unit of grouping both methods would seem to give comparable results. The differences between the estimates made by the same method are much greater, and the effect

of increasing the unit of grouping, and in this way shortening the labour, is to increase the value of r quite appreciably. Whether, or not, we should regard these estimates for the three months age grouping as satisfactory would depend on the degree of accuracy required in any particular calculation. It must be remembered, however, that the basic numerical data are of rather an extreme type in this example. It is doubtful whether any naturally living rat population would have so good a life table and so high a degree of fertility as that assumed for this imaginary population. In fact it was for these very reasons that these data were

	$L_x m_x$ column	Matrix	Difference
1 month age groups	0·44565	0·44626	0·00061
3 ,,	0·46034	0·4634	0·0031
Difference	0·01469	0·0171	

chosen as the basis of the numerical calculations in this work. For, if it could be shown that the two methods of computation gave comparable results in this case, it was felt that an even better agreement should be obtained in less extreme examples and more particularly with data relating to populations whose rate of increase is nearer to the stationary state. Although in this example the larger unit of grouping leads to rather unsatisfactory estimates of the rate of increase and the stable age distribution, it seems probable that, for the reasons just given, the differences would be less for instance in the case of human data. Hence, the question of the unit to be adopted is likely to become of less importance in the type of data more commonly met with, though it would be necessary to work out an example for such a population in order to check this point.

There is, however, one way to avoid this difficulty of the working unit for populations with a high relative rate of increase. For example, returning to the numerical data used here, supposing that it was necessary in some particular problem to have a fairly high degree of accuracy in the results, but that the work involved in manipulating the large order matrix of 21×21 was too excessive. It might be sufficient in the case we are imagining to know the age distribution of the population in three month age groups at some particular time in the future, which we will take to be a multiple of three. Then, once the real latent root (λ_1) of the large matrix and its associated stable vector have been determined, it is possible to construct a small order matrix of 7×7 which has λ_1^3 as its dominant root and therefore the same real stable age distribution as the larger matrix, only expressed in three month instead of in one month age units. It is convenient to carry out the calculation in terms of the canonical form B and of ψ vectors. Having determined the dominant latent root and the stable vector for the larger matrix, the elements of the first three rows of B^3 are then written down and summed in columns. This can be done very quickly in the present example, where reproduction does not start until the age of 3 months, for the third row of B^3 is the same as the first row of B; the second row is merely the first row of B shifted one age group to the left; and similarly again for the first row. The sums of the columns are then weighted with the number alive in the appropriate age group in the stable population (ψ_1 vector), and by summing the weighted column totals in groups of three and taking the weighted mean, we obtain the elements of the first row of a 7×7 matrix which has λ_1^3 as its domi-

nant latent root. Thus, the characteristic equation of the original matrix condensed in this way was

$$\lambda^7 - 1{\cdot}5056\lambda^6 - 6{\cdot}4694\lambda^5 - 7{\cdot}2047\lambda^4 - 5{\cdot}5371\lambda^3 - 3{\cdot}4537\lambda^2 - 1{\cdot}3451\lambda - 0{\cdot}1447 = 0,$$

and, out of idle curiosity, all the seven roots were extracted in order to compare them with those of the previous example of a condensed matrix. The estimation in this case was carried out to a higher degree of accuracy. The results were:

$$\begin{aligned}
\lambda_1 &= 3{\cdot}81452,\\
\lambda_2 &= -1{\cdot}02526,\\
\lambda_3\lambda_4 &= -0{\cdot}5905 \pm 0{\cdot}3782i \quad (\text{mod.} = 0{\cdot}70125),\\
\lambda_5\lambda_6 &= 0{\cdot}0280 \pm 0{\cdot}6879i \quad (\text{mod.} = 0{\cdot}68847),\\
\lambda_7 &= -0{\cdot}15876.
\end{aligned}$$

The dominant root of the original matrix was 1·56246 and the cube root of λ_1 is 1·56248. The remaining roots may be compared with those given for the previous example. The two negative roots are very similar and, in the second case, the two pairs of complex roots appear to have changed places, the real part of one pair becoming positive instead of negative. Although the cube root of λ_1 is equal to the dominant root of the original matrix, it is unfortunately not true that a similar relationship holds for the remaining six values of λ. There is, for instance, no negative latent root > 1 for the larger form in this actual example.

This point, however, raises an extremely interesting question. For a given series of data a finite matrix of a relatively small order may be constructed, as in the first example given here of a condensed matrix. Supposing that the order of this matrix is increased step by step and that in each case the latent roots are found. Then, in this approach to an infinite matrix, how do the latent roots behave and what relation does the array of roots in each case bear to those of the preceding steps? For the purposes of comparison it will be necessary to express the roots in terms of some suitable unit of time, e.g. per month or per year. So far as the real positive root is concerned, it seems likely that the series of individual roots will approach nearer and nearer to a limiting value. For the root λ_1 is the ratio N_{t+h}/N_t, or the number of times the stable population has increased at the end of the interval of time h. Then, expressing λ_1 in the chosen unit of time, we have $\Lambda_1 = (\lambda_1)^{1/h}$, and taking logarithms;

$$\log_e \Lambda_1 = \frac{\log_e N_{t+h} - \log_e N_t}{h},$$

so that, when the interval of time becomes very small, corresponding to a matrix of a very large order, and $h \to 0$, the right-hand side of this equation approaches the limit $\frac{1}{N}\frac{dN}{dt} = \rho$, the true instantaneous relative rate of increase of the stable population. This argument is put forward with a certain amount of diffidence; it is only too easy for the biologist to overlook some flaw which will be immediately obvious to the trained mathematician. But, even if it were a valid argument for the behaviour of the dominant root, it can hardly be extended in this form to the case of the remaining roots; and thus the main question is left unanswered. From the point of view of the biologist, it would be interesting to know whether with an increase in the size of the matrix the array of secondary roots tends to coalesce round certain values of $\lambda^{1/h}$.

16. FURTHER PRACTICAL APPLICATIONS

If we wish merely to estimate the inherent rate of increase and the stable age distribution appropriate to some system of age specific fertility and mortality rates, there is evidently little to choose between the matrix and the ordinary methods of computation. The advantages of expressing the basic rates in the form of a matrix are more clearly seen in considering the type of problem such as the following. Let us suppose that a species of mammal at a certain season of the year invades a fresh environment where there is an ample food supply, a freedom from predators, and plenty of space to accommodate any rapid increase in numbers which might take place. Under these conditions it might be assumed for theoretical purposes that some age specific rates of fertility and mortality would remain approximately constant over a period of time. The age distribution of these immigrants then becomes of some importance owing to the effect which it must necessarily have on the future course of events. For this initial distribution must clearly be very different from that which would ultimately be established in the case of a species, such as a rodent, with possibly a very rapid rate of increase, since nestlings will not be represented in it and young individuals may be present in only relatively small numbers. Supposing then that we have a number of such populations subject to the same age schedules of fertility and mortality, but differing in the age distribution of the original immigrants, we may have reason to enquire how far the development of these populations is affected over a limited period of time by the varying form of this initial distribution, assuming for simplicity that no further waves of immigration occur.

If an estimate of the number and age distribution of the female population at successive intervals of time is alone required, the answer for any form of initial distribution is readily obtained once the series of matrices M, M^2, M^3, ..., M^t have been constructed. But, in addition, we may require to know the changes which might be expected to occur in the birth rate and death rate, and also, for example, in some such rate as the percentage of adult females pregnant, a figure which is one of the simplest measures we have of the degree of fertility among wild populations. Again, in a species like the wild rat we never know the exact age of individuals caught in the field, and thus the only measure of the form of the female age distribution is the percentage of immature females, provided, of course, we are sampling the complete population. Some method is therefore required for calculating such rates at successive intervals of time.

Once the age distribution of the female population at time t is known, an estimate of the expected number of female births per unit of time may be obtained by operating on the age distribution with the maternal frequency figures. Thus, in matrix notation we may write, the number of female births equals $m_x M^t \xi_0$, where ξ_0 is the initial age distribution and the m_x figures are treated as a row vector. Similarly the estimated number of deaths per unit of time may be obtained with the help of the age specific death rates (D_x). The relative rate of increase calculated in this way is not necessarily exact, but it may be sufficiently accurate for our present purposes. As an example of the degree of error involved in this method, we may compare the values of the stable birth rate and death rate, as given in the appendix, with those derived from the matrix stable distribution in Table 2 by operating with the m_x and D_x figures. The latter were in this case computed from the stationary age distribution and the d_x column of the life table ($D_x = d_x/L_x$). The results were as follows:

	'True' values from appendix	By operating with m_x and D_x on matrix stable distribution
Birth-rate (b)	0·51265	0·51257
Death-rate (d)	0·06700	0·06154
$b-d=r$	0·44565	0·45103

The rate of increase is overestimated by about 5·4 per thousand per month, the principal error being in the death rate. This discrepancy is due to the fact that the number of deaths under 2 months of age is underestimated by applying age specific death rates, which are based on the stationary age distribution, to the stable population grouped in one month intervals at these ages. The difference between these distributions happens to be quite marked in this example. The degree of error, however, is not very great; and in the type of problem we are considering, when the age distribution of a population may take any form, this seems to be the only practical method of estimating the rate of increase.*

Supposing, then that in the case of the rat population used here as a numerical illustration, we wished to estimate the number of females, the birth rate and death rate, and the percentage of immature females at monthly intervals up to—say—7 months from the origin of the time scale, when the initial immigration is assumed to take place. Since the jth column of the matrix gives the age distribution of the survivors and the surviving descendants per individual female alive in the age group $j-1$ to j at $t=0$, the sum of the elements in this column gives the number of times the original population in this age group has increased, or decreased, at time t. The percentage of immature females may be obtained once the sum of the first three elements in the column is known (reproduction begins at the age of 3 months in this example); and the number of births and deaths per unit of time may be found by operating on the column with the m_x and D_x figures. Each of these totals, of course, will have to be multiplied in the end by the number of females alive in this age group at $t=0$. Since the initial age distribution may be of any form under the conditions of the problem, it will be necessary first of all to calculate these four totals for every column of each of the seven matrices M^t. Now, to add up the elements forming each column of a matrix is equivalent to premultiplying the matrix by a row vector of units; the sum of the first three elements may be obtained by premultiplying with a row vector of which the first three elements are units and the remainder zeros; and similarly the numbers of births and deaths are found with the help of the row vectors m_x and D_x. Thus, the operations which it is necessary to perform on

* Another similar method, which avoids the actual calculation of the number of deaths by means of the age specific death rates, is suggested by the following relationship. If the transformed age distribution $\psi = H\xi$, where H is the matrix defined in § 5, is operated on with a row vector which consists of the $L_x m_x$ figures (Appendix, Table 4), and the resulting scalar is divided by the sum of the elements of ψ, an estimate is obtained of the relative rate of increase of a population with an age distribution ξ in the original co-ordinate system. This follows from the properties of the transformed population discussed at the end of § 5 and from the relationship between the first row of the canonical form $B = HAH^{-1}$ and the $L_x m_x$ column (§ 6). In the transformed population the death rate $=0$, and the maternal frequency is given by $L_x m_x$. Thus, by transforming back again to the original co-ordinate system

$$r = \frac{[L_x m_x]\, H\xi}{[1]\, H\xi},$$

where [1] defines a row vector of units. Taking ξ as the matrix stable distribution of Table 2, and calculating the row vectors $[L_x m_x]\, H$ and $[1]\, H$, the value of r was estimated by this method to be 0·44468.

each of the columns of M^t may be written as the matrix R, which will consist of $m+1$ columns and n rows, the number of the latter depending on the number of operations. Then the required totals for the matrix M^t will be given by

$$Z^t = RM^t = RMMM \dots M,$$

and it is easy to see how the Z matrices may be built up in succession without calculating the actual matrices M^2, M^3, M^4, etc. Once the series of Z matrices have been constructed, we can obtain from $Z^t\xi_0$ the necessary figures from which the required rates at time t for a population with an initial age distribution ξ_0 may be calculated. Moreover, if we wish, the contributions made, for instance, to the total number of births or deaths by any particular age group in the initial distribution can also be determined.

The computations in this illustration have been greatly simplified by the assumption that the system of age specific fertility and mortality rates remains constant. In the case when the basic matrix M is changing with time and the age distribution at time t is given by $M_t \dots M_3 M_2 M_1 \xi_0$, some of the rows of R will also be varying. Hence the series of Z matrices could not be built up without first computing M_2M_1, $M_3M_2M_1$, etc. The latter, however, may often be of interest in themselves. For, if each column of M^t—or $M_t \dots M_3 M_2 M_1$ in the case of a variable matrix—is multiplied by the number of females alive in the appropriate age group at $t = 0$, the complete age structure of the population at time t is represented in the form of a two-dimensional array. Since the sum of the elements in each row is the total number alive in the age group x to $x+1$ at time t, the number contributed to this total by each age group at $t = 0$ is given by the individual entries.

APPENDIX

(1) *The tables of mortality and fertility*

The basic life table and fertility table which have been used in the numerical part of this study are given in Table 4. The adult l_x figures from the age of 2 months onwards are based on the mortality observed among the females of a domesticated brown rat stock housed at the Wistar Institute, Philadelphia. According to the data for 26 generations of this laboratory stock published by King (1939) it appears that out of 1384 females alive at the age of 2 months (60 days), 1337 were alive at 12 months, and 984 at 20 months. This information gave three points on the l_x curve, supposing that these survivors could be regarded as ordinates at these exact ages. In order to interpolate for other ages, a logistic type of curve was fitted to the data, the values of the constants being chosen so that the curve passed through these three points. The l_x values in Table 4 are given by

$$l_x = \frac{0{\cdot}85156355}{1+0{\cdot}00101065e^{0{\cdot}30016x}}, \quad \text{for} \quad x \geqq 2.$$

Although the original data did not extend beyond the age of 20 months, by which time the vast majority of the females had ceased breeding (King, 1939), this l_x curve was extrapolated to later ages, whenever necessary, simply for the purposes of this theoretical study.

The degree of infant mortality assumed here, namely 15 % between birth and the age of 2 months, is entirely arbitrary; it represents a moderate degree of loss at these early ages. Some care, however, was taken to weld the infant mortality smoothly on to the remainder of the l_x curve, and it was assumed that the number of deaths according to age (d_x) decreased

geometrically between birth and the age of 2 months. The actual calculations for these age groups were carried out in units of 1/8 of a month and the resulting l_x curve was integrated by means of Simpson's rule. The same method of numerical integration was used also for the adult part of the life table in order to obtain the L_x figures.

The fertility table is partly artificial and was constructed in the following way. The gross reproduction rate of these domesticated brown rats was estimated from the data published by King (1939) to have been just under 10 litters for the later generations, when the stock was thoroughly adapted to life in the laboratory. The frequency of litter production according to the age of the mother has been found by the author (unpublished observations) to be represented closely by a Pearsonian type I curve in the case of certain litter fertility tables, for example in the cross-albino rat, the vole and some human populations with a high degree of fertility; and, moreover, the values of β_1 and β_2 were very similar for all three species. The actual equation for the curve used here may be written

$$y = y' x^{1\cdot5-1}(a-x)^{2\cdot5-1},$$

and the range was assumed to be from 3 to 21 months, which for grouping purposes represents the span of the reproductive ages observed by King in this Wistar strain of brown rats. The ordinates of the integral curve of the above equation were taken from the tables of the incomplete Beta function and, in this way, a column was formed which gave a gross reproduction rate of 10 litters. The individual entries were then multiplied by the mean number of daughters per litter according to the age of the mother, which was recorded by King, and thus the m_x figures in Table 4 were obtained. The gross reproduction rate is 31·21 daughters and the net rate 25·66. These tables of fertility and mortality were originally constructed in order to determine the relative rate of increase and the type of stable age distribution which might be expected in a brown rat population living under more or less optimum conditions.

(2) *Calculation of the rate of increase*

Some difficulty was experienced in obtaining a satisfactory estimate of the rate of increase (r) from the usual solution (Dublin & Lotka, 1925) of the equation:

$$\int_0^\infty e^{-rx} l_x m_x dx = 1.$$

The 4th degree equation in r with the numerical coefficients based on the seminvariants of the $L_x m_x$ distribution, the estimates of which are given in § 6, was

$$4{\cdot}90199r^4 + 3{\cdot}69283r^3 - 7{\cdot}07198r^2 + 9{\cdot}60604r - 3{\cdot}2448498 = 0,$$

and the real root was found to be 0·42447. This value of r was, however, clearly too low and a better estimate had to be obtained in a rather roundabout way, since it was thought that the use of higher moments than the fourth would be unsatisfactory in the present example.

If the force of mortality represented by the original life table is increased by a constant factor (r) which is independent of age, the new life table is

$$l'_x = e^{-rx} l_x,$$

and the net reproduction rate will be given by $R_r = \Sigma L'_x m_x$, where L'_x are the integrals of the new l'_x curve. Clearly, the greater r is taken to be, the smaller R_r becomes. Then, suppose that the relation between R_r and r is given by

$$\log_e R_r = a + br + cr^2 + dr^3 + \ldots,$$

Table 4

Age (x) Units of 30 days	Life table		Fertility table	
	l_x	L_x	m_x	$L_x m_x$
0	1·00000	0·46544	—	—
0·5	0·88706	0·43489	—	—
1	0·85882	0·42725	—	—
1·5	0·85176	0·42534	—	—
2	0·85000	0·84973	—	—
3	0·84945	0·84910	1·1342	0·96305
4	0·84871	0·84824	2·0797	1·76408
5	0·84772	0·84708	2·6596	2·25289
6	0·84638	0·84553	2·8690	2·42582
7	0·84458	0·84344	2·9692	2·50434
8	0·84217	0·84063	2·9535	2·48280
9	0·83893	0·83687	2·8143	2·35520
10	0·83459	0·83184	2·6114	2·17227
11	0·82881	0·82515	2·2455	1·85287
12	0·82113	0·81629	2·0533	1·67609
13	0·81098	0·80463	1·7971	1·44600
14	0·79768	0·78940	1·5561	1·22839
15	0·78039	0·76975	1·2175	0·93717
16	0·75821	0·74472	0·9548	0·71106
17	0·73018	0·71342	0·6610	0·47157
18	0·69548	0·67512	0·4043	0·27295
19	0·65354	0·62953	0·1846	0·11621
20	0·60434	0·57696	0·0435	0·02510
21	0·54859	—	—	—
Total	—	—	31·2086	25·65786

where $a = \log_e R_0$, and the constants b, c, d, etc. are to be determined. It was assumed in the present instance that a 4th degree polynomial in r would be sufficient, and four new life tables were constructed taking r to be in turn 0·1, 0·2, 0·4 and 0·5. The L'_x integrals were obtained by Simpson's rule for the reproductive ages and the four values of R_r calculated. The equations for finding the values of the constants were:

$$0{\cdot}0001e + 0{\cdot}001d + 0{\cdot}01c + 0{\cdot}1b = -0{\cdot}8934974,$$
$$0{\cdot}0016e + 0{\cdot}008d + 0{\cdot}04c + 0{\cdot}2b = -1{\cdot}6708610,$$
$$0{\cdot}0256e + 0{\cdot}064d + 0{\cdot}16c + 0{\cdot}4b = -2{\cdot}9792984,$$
$$0{\cdot}0625e + 0{\cdot}125d + 0{\cdot}25c + 0{\cdot}5b = -3{\cdot}5490542,$$

whence

$$b = -9{\cdot}617235,$$
$$c = \ \ 7{\cdot}371816,$$
$$d = -5{\cdot}698291,$$
$$e = \ \ 2{\cdot}062332.$$

Inserting these values of the constants in the equation for $\log_e R_r$, the value of r for which $\log_e R_r = 0$ was found to be 0·44565.

The stable birth rate was estimated in the usual way from

$$\frac{1}{b} = \int_0^\infty e^{-rx} l_x \, dx.$$

The integrals were computed by means of Simpson's rule, treating the age groups 0–2 separately from the remainder of the life table, the units adopted being 1/4 month for the early ages compared with 1 month for the later. It was found that

$$\int_0^{21} e^{-0\cdot 44565x} l_x dx = 1\cdot 95064,$$

$$b = 0\cdot 51265,$$

and hence the death rate,

$$d = 0\cdot 06700.$$

The value of r is so high in this case that the error in the estimate of b due to neglecting the ages from 21 onwards would only be in the last figure. The stable age distribution is given in Tables 1 and 2 of the text. Owing to the way in which the value of r was determined, it will be found that the birth rate of the stable population obtained by operating on this age distribution with the maternal frequency figures is precisely the same as that given by the above integral.

(3) *Numerical values of the matrix elements*

The numerical elements of the matrices A and B to which reference has been made in §§ 3 and 6 of the text are given in Table 5.

Table 5

Age group x	Matrix A		Matrix B
	P_x	F_x	Elements of first row
0–	0·94697	0	0
1–	0·99665	0	0
2–	0·99926	0·3964	0·3741
3–	0·99899	1·4939	1·4089
4–	0·99863	2·1777	2·0517
5–	0·99817	2·5250	2·3756
6–	0·99753	2·6282	2·4682
7–	0·99667	2·6749	2·5059
8–	0·99553	2·6018	2·4293
9–	0·99399	2·4419	2·2698
10–	0·99196	2·1865	2·0202
11–	0·98926	1·9044	1·7454
12–	0·98572	1·7259	1·5648
13–	0·98107	1·4918	1·3332
14–	0·97511	1·2415	1·0885
15–	0·96748	0·9522	0·8141
16–	0·95797	0·7141	0·5907
17–	0·94631	0·4618	0·3659
18–	0·93247	0·2518	0·1888
19–	0·91649	0·0901	0·0630
20–		0·0035	0·0022

I wish to thank Professor M. Greenwood, F.R.S. and Dr J. O. Irwin for their kindness in reading the manuscript, and also Dr H. Motz for many fruitful discussions. The study arose out of some research work which is being carried out by the Bureau of Animal Population with the aid of a grant from the Agricultural Research Council.

REFERENCES

Aitken, A. C. (1939). *Determinants and Matrices*, 1st ed. Edinburgh: Oliver and Boyd.

Bernardelli, H. (1941). Population waves. *J. Burma Res. Soc.* **31**, no. 1, 1–18.

Chapman, R. N. (1933). The causes of fluctuations of populations of insects. *Proc. Hawaii Ent. Soc.* **8**, no. 2, 279–92.

Charles, Enid (1938). The effect of present trends in fertility and mortality upon the future population of Great Britain and upon its age composition. Chap. II in *Political Arithmetic*, ed. Lancelot Hogben. London: Allen and Unwin.

Dublin, L. I. & Lotka, A. J. (1925). On the true rate of natural increase as exemplified by the population of the United States, 1920. *J. Amer. Statist. Ass.* **20**, 305–39.

Frazer, R. A., Duncan, W. J. & Collar, A. R. (1938). *Elementary Matrices*. Camb. Univ. Press.

Glass, D. V. (1940). *Population Policies and Movements in Europe*. Oxford: Clarendon Press.

King, Helen D. (1939). Life processes in grey Norway rats during fourteen years in captivity. *Amer. Anat. Mem.* no. 17, 1–72.

Turnbull, H. W. & Aitken, A. C. (1932). *An Introduction to the Theory of Canonical Matrices*. London and Glasgow: Blackie and Son.

Whittaker, E. T. & Robinson, G. (1932). *The Calculus of Observations*, 2nd ed. London and Glasgow: Blackie and Son.

ERRORS IN THE ROUTINE DAILY MEASUREMENT OF THE PUERPERAL UTERUS

By C. SCOTT RUSSELL, M.A., M.B., Ch.B., F.R.C.S.Ed., M.R.C.O.G.
First Assistant, Nuffield Department of Obstetrics and Gynaecology, University of Oxford

In both district and institutional midwifery, it is usual to measure each day the height of the puerperal uterus above the symphysis pubis. Though most observers would say that the purpose behind this custom is the detection of uterine subinvolution, there is no agreement as to the precise conditions that must be fulfilled or the criteria required for establishing such a diagnosis. The present paper is concerned with the systematic and the uncontrolled errors of this routine measurement, and with the significance of fluctuations in the observed height of the puerperal uterus.

In institutions, the puerperal uterus is usually measured by whichever member of the nursing staff happens to be available: often the measurements are made by pupil midwives. It must be rare for a uterus in any period of 7 days to be measured by only one observer. The measurement, which is made between the upper surface of the uterus and the upper surface of the symphysis pubis, is recorded on the temperature chart; it is probably never recorded more accurately than to the nearest $\frac{1}{4}$ in., and usually the nearest $\frac{1}{2}$ in. is considered sufficiently accurate.

Normal uterine involution

It is widely taught that the normal human uterus involutes $\frac{1}{2}$ in. a day during the puerperium. This is only an approximation; if it were true, the puerperal uterus having shrunk in 1 day the last $\frac{1}{2}$ in., would suddenly stop involuting. This would be contrary to clinical experience which tells us that the process starts fast and finishes slowly. The changing daily amount of normal involution would be too cumbersome a standard in practice; besides, great accuracy is not necessary when the uterus is measured only once a day and slight irregularities are of little clinical importance. A rate of $\frac{1}{2}$ in. a day may therefore be taken as accurate enough for practical purposes.

Factors influencing the recorded height of the uterus

The factors that influence the recorded height of the puerperal uterus may be divided into two broad groups. First, there are those which act constantly in one direction, for example, the distended bladder or lower bowel. Secondly, there are the uncontrolled factors, such as the thickness of the abdominal wall, the mobility of the uterus, the uterine tone, the method of measurement, the points chosen for measuring, the individual differences between observers and so on. Though these many factors are just as likely to cause the reading to be too high as too low, they are important because together they constitute the error of the measurement; without knowledge of this error, logical deductions based on the measurements are not possible. The following five studies were made to obtain estimates of this error.

Method

First investigation

Seven puerperal patients whose uteri could be felt were chosen, and six unpractised observers measured each uterus twice, the interval between measurements being about

15 min. For convenience, medical students were used as observers; it was thought that they would be neither more nor less accurate than pupil midwives. An effort was made to prevent conscious or subconscious bias on the part of each observer (1) by explaining the nature of

Table 1. *Duplicate measurements of puerperal uterus by unpractised observers*

Observers' duplicate readings	Patients							Observers' totals	
	1	2	3	4	5	6	7		
1	5·00 5·50	5·00 4·25	4·25 4·75	4·00 4·50	4·25 4·75	5·75 6·00	4·75 5·50	33·00 35·25	68·25
2	4·75 5·00	4·75 4·25	4·25 3·75	4·75 4·00	4·50 4·25	5·50 5·25	5·00 5·00	33·50 31·50	65·00
3	5·50 5·75	4·75 4·75	4·00 4·50	4·25 4·50	4·75 4·75	5·75 5·75	5·00 5·50	34·00 35·50	69·50
4	6·00 6·75	5·00 5·75	4·50 5·25	3·75 4·50	5·50 6·50	5·50 6·00	5·50 5·75	35·75 40·50	76·25
5	5·25 5·75	4·50 4·75	4·00 5·00	4·00 4·50	5·75 5·25	6·00 5·75	4·50 5·25	34·00 36·25	70·25
6	5·25 6·00	4·50 5·00	4·25 5·00	4·25 4·75	6·00 5·75	5·50 5·75	5·25 4·75	35·00 37·00	72·00
Patients' totals	31·75 34·75	28·50 28·75	25·25 28·25	25·00 26·75	30·75 31·25	34·00 34·50	30·00 31·75	205·25 216·00	
	66·50	57·25	53·50	51·75	62·00	68·50	61·75		421·25

Analysis of variance

	Sum of squares	Degrees of freedom	Mean square	Variance ratio	Probability
Main effects					
Patients	19·9349	6	3·3225	43·6310	Less than 0·001
Observers	5·0845	5	1·0169	13·3539	Less than 0·001
Readings	1·3778	1	1·3778	18·0932	Less than 0·001
First order interactions					
Patients-observers	5·6187	30	0·1873	2·4576	0·05–0·01
Patients-readings	0·6798	6	0·1133	1·4879	Greater than 0·20
Observers-readings	1·6891	5	0·3378	4·4360	Less than 0·01
Second order interactions					
Patients-observers-readings [error]	2·2845	30	0·0762		
Total	36·6693	83			

the investigation, and (2) by ensuring that previous readings of other observers were not known until subsequent readings were recorded. Measurements to the nearest $\frac{1}{4}$ in. were requested. In Table 1 the measurements are recorded together with the corresponding analysis of variance. This analysis and certain important facts arising from it must be considered in detail.

Analysis of variance

The total variability of the eighty-four separate observations may be expressed by a 'mean square', i.e. the sum of squares of deviations of each observation from the arithmetic mean divided by the degrees of freedom. This variability, as expressed by a 'mean square', is divisible into mean squares due to the main effects and the first order interactions together with a remainder which is the best first estimate of uncontrolled error. Based on 30 degrees of freedom this has a mean square of 0·076, against which all the other mean squares are compared to give the different variance ratios. From these the probabilities are found from the appropriate tables (Fisher & Yates, 1943).

The interaction 'patients-readings' as expressed by the corresponding variance ratio is not significant;* the interval between readings, therefore, appeared to have had a similar effect on the seven patients. The interaction 'observers-readings' is highly significant.* This is found, by an analysis of variance on the first and second readings taken separately, to be entirely due to the second readings. Whereas the variance ratio 'between observers' obtained from the first readings is not significant (0·933), that from the second readings is highly significant (11·03). A possible explanation is that the observers largely responsible subconsciously altered the technique of measurement between reading 1 and reading 2. The interaction 'patients-observers' is significant,* suggesting that the observers found the different uteri difficult to measure in different ways.

Passing to the main effects, the variance ratio 'patients' is highly significant, as would be expected from patients observed on different days of the puerperium. The variance ratio 'observers' is highly significant, emphasizing the importance of the personal factor. The variance ratio 'readings' is also highly significant; this may be due to an accumulation of urine between the first and second readings causing slight upward displacement of the uterus.

The analysis of variance has been taken to the limit. Of the three first order interactions, the first and third are not controllable with unpractised observers; the second is not significant. All three may therefore be included in the estimate of uncontrolled error. The mean square of the adjusted error is now 0·1447. The variance of the difference between two observations is therefore 0·2894 and the standard error of the difference, the square root of this, i.e. 0·5380. If the effect of different observers is taken into account, the mean square 'observers' must be included in the error estimate. The mean square for error is then 0·2021, and the standard error of the difference between two readings 0·6358.

Second investigation

A similar study was carried out with seven other unpractised observers and seven other patients. In addition to the duplicate readings being separated by an interval of time—approximately 15 min.—each patients was given 0·5 mg. ergometrine by mouth as soon as the first readings had been completed. The results are given in Table 2. All the main effects and all the first order interactions are highly significant, supporting the evidence already presented that the measurement not only is difficult but also is influenced by many factors. The three first order interactions would not be controlled in practice, so that all may reasonably be included in the error estimate. This addition alters the mean square for error to 0·1621. The standard error of the difference between two observations is therefore 0·5694,

* The term 'significant' is used where the probability is between 0·05 and 0·01. The term 'highly significant' is used where the probability is less than 0·01. 'Not significant' refers to a probability greater than 0·05.

which is in close agreement with the value 0·5380 obtained from the first investigation. The inclusion of the effect 'between observers' raises the standard error of the difference to 0·6821, which also agrees closely with the corresponding figure 0·6358 obtained from the first investigation.

Table 2. *Duplicate measurements by unpractised observers before and after ergometrine*

Observers' readings before and after ergometrine	Patients							Observers' totals	
	1	2	3	4	5	6	7		
1	5·25	5·50	3·25	4·75	6·25	6·50	4·50	36·00	68·00
	4·75	4·50	2·75	3·50	5·50	6·00	5·00	32·00	
2	6·00	5·50	3·00	4·25	6·50	6·50	5·00	36·75	72·25
	5·25	5·50	3·00	4·50	6·00	6·00	5·25	35·50	
3	4·50	4·75	3·00	4·50	6·00	5·75	4·75	33·25	63·50
	4·50	4·50	2·75	4·25	5·50	4·50	4·25	30·25	
4	4·50	5·50	4·25	5·00	5·75	5·50	4·75	35·25	66·00
	4·50	5·00	3·00	4·25	5·00	4·50	4·50	30·75	
5	5·00	5·50	2·75	4·75	6·00	6·50	5·50	36·00	72·75
	5·00	5·50	3·00	5·00	6·50	6·00	5·75	36·75	
6	4·50	5·25	2·75	4·50	6·00	5·75	4·25	33·00	64·75
	4·75	5·00	2·25	4·75	5·75	4·75	4·50	31·75	
7	5·50	5·25	3·50	4·00	6·25	6·25	5·25	36·00	73·50
	5·75	5·75	3·50	5·00	6·25	5·75	5·50	37·50	
Patients' totals	35·25	37·25	22·50	31·75	42·75	42·75	34·00	246·25	
	34·50	35·75	20·25	31·25	40·50	37·50	34·75	234·50	
	69·75	73·00	42·75	63·00	83·25	80·25	68·75		480·75

Analysis of variance

	Sum of squares	Degrees of freedom	Mean square	Variance ratio	Probability
Between patients	76·4707	6	12·7451	182·0707	Less than 0·001
Between observers	7·3190	6	1·2198	17·4257	Less than 0·001
Between readings	1·4089	1	1·4089	20·1271	Less than 0·001
Patients-observers	7·3061	36	0·2029	2·8986	Less than 0·01
Patients-readings	1·5420	6	0·2570	3·6714	Less than 0·01
Observers-readings	2·2473	6	0·3745	5·3500	Less than 0·001
Error	2·5205	36	0·0700		
Totals	98·8145	97			

Error of practised observers (third investigation)

The error of measurement with unpractised observers was of special importance for those institutions where the routine puerperal uterine measurements are made by pupil midwives. Elsewhere the routine is the responsibility of comparatively senior nurses. In order to

obtain an estimate of the error of trained observers another simple study was made. Ten puerperal patients each had the uterus measured by five observers of whom four were nurses and one was a doctor. The observers were experienced in the measurement; they were asked to use the method to which they were accustomed. Each observer measured each uterus once. In Table 3 the measurements are tabulated with the corresponding analysis of variance.

It should be noted that not one of the measurements was given more accurately than to the nearest ¼ in., and that forty-four out of the fifty readings were given to the nearest ½ in.

The two main effects, 'between patients' and 'between observers', are both highly significant, as would be expected. The remainder or error estimate has a mean square of 0·3308, which is considerably higher than that of the unpractised observers. This apparent paradox

Table 3. *Measurements of the puerperal uterus by practised observers*

Observers	Patients										Observers' totals
	1	2	3	4	5	6	7	8	9	10	
1	5·00	5·00	5·00	3·00	4·00	4·50	2·50	4·00	4·00	6·25	43·25
2	4·25	4·00	5·75	1·50	2·50	3·00	1·00	2·50	3·00	5·00	32·50
3	4·50	4·50	5·00	2·00	2·50	2·00	3·00	1·50	4·00	6·00	35·00
4	4·50	5·00	6·00	2·50	4·50	3·00	1·50	4·00	3·50	5·50	40·00
5	4·50	5·00	5·50	3·00	4·00	4·25	2·25	3·50	4·25	6·50	42·75
Patients' totals	22·75	23·50	27·25	12·00	17·50	16·75	10·25	15·50	18·75	29·25	193·50

Analysis of variance

	Sum of squares	Degrees of freedom	Mean square	Variance ratio	Probability
Between patients	70·2800	9	7·8088	23·608	Less than 0·001
Between observers	9·0925	4	2·2731	6·87	Less than 0·001
Error	11·9075	36	0·3308		
Total	91·2800	49			

is probably largely due to the fact that the unpractised observers were asked to measure to the nearest ¼ in. and thereby encouraged to be accurate, whereas four out of five of the practised observers made the measurement to the nearest ½ in.—except on three occasions —as this was their usual practice.

Error of individual observers (investigations 4 and 5)

Two studies were now made to obtain estimates of the error of individual persons.

An experienced midwifery sister tutor made five replicate readings to the nearest ¼ in. on nine puerperal patients. The results are shown in Table 4. The mean square between readings was not significant and was therefore included in the error estimate. The mean square for error was now 0·1639, which is not appreciably different from 0·1447 and 0·1621 which are the corresponding figures previously obtained.

Having had special experience of the measurement, I repeated the above study. Of twelve unselected puerperal patients only eight were measured, the remaining four being thought unsuitable because of abdominal tenderness (one case), recent delivery (one case), and difficulty in palpating the uterus (two cases). This partial selection of cases left for analysis only those whose uteri could be felt fairly easily. The results are shown in Table 5. The mean square between readings was again not significant and was included in the estimate of error. The new value was 0·0813, which is lower than the previous best. The improvement, though partly the result of greater care in measurement, is probably largely due to some selection of material.

Table 4. *Replicate measurements by midwifery sister tutor*

Replicate readings	Patients									Readings' totals
	1	2	3	4	5	6	7	8	9	
1	5·00	5·75	3·00	2·00	3·75	3·00	4·00	4·50	6·25	37·25
2	5·00	6·00	2·75	2·75	3·75	3·00	5·00	4·50	6·00	38·75
3	5·00	5·00	3·00	3·00	4·75	3·00	4·75	4·50	6·00	39·00
4	4·75	6·00	3·50	4·00	4·75	3·25	4·50	4·25	6·00	41·00
5	4·50	5·00	3·00	3·75	4·50	3·00	4·75	4·50	6·00	39·00
Patients' totals	24·25	27·75	15·25	15·50	21·50	15·25	23·00	22·25	30·25	195·00

Analysis of variance

	Sum of squares	Degrees of freedom	Mean square	Variance ratio	Probability
Between patients	47·9750	8	5·9969	37·565	Less than 0·001
Between readings	0·7916	4	0·1979	1·240	Greater than 0·20
Error	5·1084	32	0·1596		
Readings + error	5·9000	36	0·1639		
Totals	53·8750	44			

Effect of the bladder on the puerperal uterus

Attention must now be directed to the factors which bias the recorded height of the uterus. The only factor which constantly makes the recorded size smaller than reality is a tendency for the uterus to fall backward into the pelvis towards the end of the first week of the puerperium. There are, however, two important influences tending to make the uterus seem larger than reality. The one is the urinary bladder, of special importance in the immediate postpartum period: the other is the pelvic colon and rectum, of more importance a few days later The effect of the bladder is well known, and I have even found the uterine fundus displaced so far upwards by the bladder that it could be felt under the costal margin and ballotted like the foetal head in a breech presentation. To show this upwards displacement the following short study was made.

Table 5. *Replicate measurements by author*

Replicate observations	Patients								Readings' totals
	1	2	3	4	5	6	7	8	
1	4·75	3·50	2·50	3·00	5·00	5·25	7·00	4·25	35·25
2	4·25	3·25	2·00	3·00	5·25	5·25	6·50	4·75	34·25
3	4·50	3·50	2·50	3·00	5·50	5·25	6·00	4·50	34·75
4	4·50	3·50	2·00	3·25	5·00	4·75	7·00	4·50	34·50
5	3·75	3·25	2·25	3·50	4·75	4·75	6·50	4·75	33·50
Patients' totals	21·75	17·00	11·25	15·75	25·50	25·25	33·00	22·75	172·25

Analysis of variance

	Sum of squares	Degrees of freedom	Mean square	Variance ratio	Probability
Between patients	64·4609	7	9·2087	107·83	Less than 0·001
Between readings	0·2093	4	0·0523		
Patients-readings [error]	2·3907	28	0·0854		
Readings + error	2·6000	32	0·0813		
Totals	67·0609	39			

Fourteen unselected patients were used. Each patient's uterus was measured by myself just before a morning bed-pan round. Following the use of the bed pan each uterus was again measured. In Table 6 the duplicate readings and the differences are tabulated. Being concerned with the displacement of the uterus by the bladder the appropriate statistical test is based on the difference column, in which it should be noted all fourteen figures bear the same sign. Using the t test (Fisher, 1941, p. 117), the probability is found to be less than

Table 6. *Measurements of the puerperal uterus before and after micturition*

Measurements		Difference [y]	Measurements		Difference [y]
First	Second		First	Second	
7·50	6·25	1·25	6·25	5·25	1·00
4·25	3·75	0·50	5·00	2·00	3·00
7·00	4·50	2·50	6·75	5·00	1·75
6·00	4·50	1·50	7·50	5·25	2·25
8·00	4·25	3·75	4·25	2·50	1·75
3·75	2·25	1·50	6·50	4·25	2·25
5·00	4·00	1·00	5·50	4·00	1·50

$$\text{Sum}\ \frac{(y-\bar{y})^2}{13} = 0{\cdot}7445 \quad \frac{0{\cdot}7445}{14} = 0{\cdot}0532 = (-0{\cdot}231)^2.$$

$$\bar{y} = 1{\cdot}8214 \quad t_{13} = 7{\cdot}8 \quad P < 0{\cdot}001.$$

0·001. As the condition of the bladder was the only important factor that varied between reading 1 and reading 2, there can be no doubt that the uterus had been displaced upwards.

Because one would expect a high correlation between the displacement of the uterus and the amount of urine passed, each difference was compared with the corresponding output of urine. The regression coefficient of fall in uterine height on urinary output is 0·046. The mean square for error is very high—further evidence of the difficulty in measuring the puerperal uterus—and the regression coefficient is, in fact, not significantly different from zero (P between 0·1 and 0·2). If further cases were studied, one would expect from clinical experience that a significant coefficient would be obtained.

Discussion

Though we cannot tell in advance precisely how much a given uterus will involute in a given time, it is known that the average involution rate of the healthy organ is, as a close approximation, $\frac{1}{2}$ in. per day during the early days of the puerperium. Therefore, the expected height of a uterus on any one of these days may be estimated as $\frac{1}{2}$ in. less than that recorded for the previous day.

The records of the uterine height are influenced by many factors outside clinical control; these are of real importance because together they constitute the error of measurement. Without knowledge of the magnitude of this error, it is impossible to judge the significance of measurements and differences between measurements.

Five investigations have been made to determine the magnitude of this error; two were with unpractised observers and three with practised observers. The standard error of the difference between two readings of a single average careful observer is approximately $\frac{1}{2}$ in. Unless, therefore, the difference between the recorded height of the uterus and the expected height is more than 1 in., i.e. twice the standard error of the difference, it is not safe to conclude that anything other than the error of the measurement has been responsible.* But, according to our estimate, the uterus should involute only $\frac{1}{2}$ in. per day. Therefore, unless the recorded height is at least $\frac{1}{2}$ in. greater than that recorded for the previous day, we should not be surprised nor assume anything to be wrong, as such a difference is likely to be due to this error only. This argument may be extended to cover comparisons between measurements at intervals up to 2 or 3 days. If taken much beyond 3 days, however, a serious fallacy might result. Our estimate of the average daily uterine involution is an approximation, and as this is multiplied, so will the error of the approximation increase.

By exercising great care, measuring to the nearest $\frac{1}{4}$ in., and selecting only those patients whose uterus can be felt easily, it should be possible to reduce the standard error of the difference between two measurements to the region of $\frac{1}{4}$ in. Any investigation designed to compare involution rates under different conditions would be helped by such increase in accuracy.

There are two common controllable influences which elevate the puerperal uterus. The one, more important in the early days of the puerperium, is a full bladder; the other, exerting its influence later, and of less importance, is a full bowel (Moir & Russell, 1943). It has now been shown, not only that the full bladder raises the uterus, but also that in

* Strictly the comparison between normal and abnormal uterine involution involves a test of significance between two regression lines. In practice, however, it would be unreasonable to expect any medical attendant to calculate the regression line on every occasion that the size of the uterus was larger than expectation. The method described should be adequate and, requiring only mental arithmetic, is suitable for the bedside.

thirteen out of fourteen patients with normal bladder distension the extent was up to or greater than twice the standard error of the difference between two measurements of uterine height. With pathological bladder distension, the upward displacement of the uterus will be still greater and correspondingly more than the error of measurement; in contrast, alteration in uterine height due to subinvolution will be less than the error of measurement.

The practical conclusion is that careful routine daily puerperal uterine measurement is of value, more because it helps in the recognition of abnormal states of the bladder or bowel than those of the uterus. Distension of the bladder or bowel should always be suspected if, from any two measurements over a period of not more than 3 or 4 days, the observed height of the uterus is more than 1 in. greater than the expected height. It must, however, be emphasized that this figure of 1 in. has been obtained from measurements carefully made to the nearest $\frac{1}{4}$ in.; it must be increased to $1\frac{1}{2}$ or even 2 in. if measurements are perfunctorily made by different persons each day. A diagnosis of uterine subinvolution is only permissible after due allowance has been made for the effect of all factors, controllable and uncontrollable, that influence the recorded height.

I wish to record my thanks for help in the preparation of this paper to Mr P. H. Leslie, Dr W. T. Russell and Professor Chassar Moir.

REFERENCES

FISHER, R. A. (1941). *Statistical Methods for Research Workers.* Oliver and Boyd.

FISHER, R. A. & YATES, F. (1943). *Statistical Tables for Biological, Agricultural and Medical Research.* Oliver and Boyd.

MOIR, J. C. & RUSSELL, C. S. (1943). *J. Obstet. Gynaec.* **50**, 94.

ON A METHOD OF ESTIMATING FREQUENCIES

By J. B. S. HALDANE, F.R.S.

In a very large variety of investigations we desire to estimate the frequency of an attribute in a series of populations, each of which is so much larger than the sample taken from it that it may be regarded as infinite. If p be the frequency of the attribute, $q = 1-p$, and n the number in the sample, the standard error of the estimate of p is of course $\sqrt{(pq/n)}$. Provided p does not vary much from one sample to another, it is desirable to keep n approximately constant. But when p varies greatly this is unsatisfactory. Thus if $n = 1000$, and $p = 0{\cdot}3$, its standard error is 0·015, but if p is 0·01 its standard error is 0·0031, so that we could not distinguish between populations where p was 0·01 and 0·005. In such a case it may be desired that the standard error of each value of p should be roughly proportional to p rather than to its square root.

Such cases arise in haematology. My friend Dr R. A. M. Case has been investigating the frequency of siderocytes, an abnormal type of red blood corpuscle described by Grüneberg (1942), in a number of bloods. Their frequency ranged from about 0·2 to 20 % or more. He has adopted the method of counting stained red corpuscles in a film until he had counted some definite number m, usually about 20, of siderocytes. If, in order to count this number, he had observed a total of n red corpuscles, he took $\dfrac{m-1}{n}$ as his estimate of p. It will be shown that the correct estimate is $\dfrac{m-1}{n-1}$, so that his error was negligible; and the standard error will be calculated.

Let p be the frequency of abnormal cells, and $q = 1-p$,
m be the number of abnormal cells counted,
n be the total number of cells counted,
$x = \dfrac{m-1}{n-1}$ be the estimate of p.

Clearly n may have any positive value exceeding $m-1$. Let w_n be the probability that exactly n cells are counted before m abnormal cells are observed. Two things are necessary and sufficient if this is to be the case. The first $n-1$ cells must include $m-1$ abnormals, and the nth cell must be abnormal. The probabilities of these two events are $\binom{n-1}{m-1} p^{m-1}q^{n-m}$, and p. Hence $w_n = \binom{n-1}{m-1} p^m q^{n-m}$. This is the coefficient of t^n in $\left(\dfrac{pt}{1-qt}\right)^m$.

The mean value of x is

$$\bar{x} = (m-1)\sum_{n=m}^{\infty} \frac{w_n}{n-1}$$

$$= \sum_{n=m}^{\infty} \binom{n-2}{m-2} p^m q^{n-m},$$

or if $n-m = r$

$$p^m \sum_{r=0}^{\infty} \binom{m+r-2}{r} q^r$$

$$= p^m(1-q)^{1-m}$$

$$= p.$$

Thus x is an unbiased estimate of p.

The modal value of x is the value given by the value of n which makes w_n maximal. $\frac{w_{n+1}}{w_n} = \frac{nq}{n-m-1}$. Hence w_n exceeds w_{n-1} and w_{n+1}, if n lies between $\frac{m-1}{p}$ and $\frac{m-1}{p}+1$. Therefore the modal value of x ranges between p and $p+\frac{p^2}{m-1-p}$. Thus the distribution of x is slightly asymmetrical, as is obvious from the fact that since n can have any value exceeding $m-1$, the range of possible values of x is from 0 to 1. If we used m/n as our estimate of p a bias would be introduced. In fact, the mean value of m/n can be shown to be

$$mp^m q^{-m}\int_0^q t^{m-1}(1-t)^{-m}\,dt, \quad \text{or} \quad p\sum_{r=0}^{\infty}\frac{m!\,r!\,q^r}{(m-r)!}.$$

On the other hand, the mean value of n/m is easily shown to be p^{-1}. Thus if we were in the habit of expressing frequencies in such a form as '1 in 20' rather than '5 %' or '0·05', the method of counting up to a fixed number m of abnormals, and dividing this into the total number n, would give us an unbiased estimate, while the method of counting the number of abnormals in a sample of fixed size n and dividing n by m would not do so. For, of course, the reciprocal of an unbiased estimate of a parameter is not an unbiased estimate of the reciprocal of the parameter.

Similarly
$$\begin{aligned}\overline{x^2} &= \sum_{n=m}^{\infty}\frac{(m-1)^2 w_n}{(n-1)^2}\\ &= (m-1)\,p^m q^{1-m}\sum_{n+m}^{\infty}\binom{n-2}{m-2}\frac{q^{n-1}}{n-1}\\ &= (m-1)\,p^m q^{1-m}\int_0^q\sum_{n=m}^{\infty}\binom{n-2}{m-2}t^{n-2}\,dt\\ &= (m-1)\,p^m q^{1-m}\int_0^q t^{m-2}(1-t)^{1-m}\,dt. \qquad (1)\end{aligned}$$

When m is small this can be integrated directly, e.g. for $m=4$,

$$\overline{x^2} = \frac{3p^2}{2q^2}\left(q-2p-\frac{2p^2\log p}{q}\right).$$

For larger values of m it is better to expand the integral in an infinite series. Put $u=\frac{pt}{q(1-t)}$, so that $t=\frac{qu}{p+qu}$, $dt=\frac{pq\,du}{(p+qu)^2}$.

Then
$$\begin{aligned}\overline{x^2} &= (m-1)\,p^2\int_0^1\frac{u^{m-2}\,du}{p+qu}\\ &= (m-1)\,p^2\int_0^1 u^{m-2}[1-q(1-u)]^{-1}\,du\\ &= (m-1)\,p^2\sum_{r=0}^{\infty}q^r\int_0^1 u^{m-2}(1-u)^r\,du\\ &= (m-1)\,p^2\sum_{r=0}^{\infty}\frac{(m-2)!\,r!\,q^r}{(m+r-1)!}\\ &= p^2\sum_{r=0}^{\infty}\frac{(m-1)!\,r!}{(m+r-1)!}q^r = p^2\sum_{r=0}^{\infty}\binom{m+r-1}{r}^{-1}q^r\\ &= p^2\left[1+\frac{q}{m}+\frac{2!\,q^2}{m(m+1)}+\frac{3!\,q^3}{m(m+1)(m+2)}+\ldots\right].\end{aligned}$$

Hence the variance of x is

$$\sigma^2 = \overline{x^2} - p^2 = \frac{p^2 q}{m}\left[1 + \frac{2!q}{m+1} + \frac{3!q^2}{(m+1)(m+2)} + \dots\right]. \quad (2)$$

If in this equation we insert the estimated values of p and q, namely, $\frac{m-1}{n-1}$ and $\frac{n-m}{n-1}$, we find

$$\sigma^2 = \frac{m(n-m)}{n^2(n-1)}\left[1 + \frac{2(n-m)(n-3m)}{m^2 n^2} + O(m^{-3})\right]. \quad (3)$$

Thus $\sigma^2 = \frac{m(n-m)}{n^2(n-1)}$ is sufficiently accurate for all purposes and the classical value $\sigma^2 = \frac{m(n-m)}{n^3}$ sufficiently for most. These values can also be taken as the variances of p when m and n are given, though of course the exact values will depend on the prior probability distribution of p. It must be noted that if the variance is calculated from the estimated value of p it is approximately $\frac{p^2 q}{m-2}$, not $\frac{p^2 q}{m}$, so long as p is small. Since σ approximates to $p\sqrt{\frac{q}{m-2}}$, the standard error of p is a nearly constant fraction of the value of p when p is small, provided m is kept constant throughout a series of determinations. The higher moments can be obtained by a similar method, but as they involve multiple integrals, the series expansions are somewhat complicated.

Suppose now that the population sampled consists of several classes, and that a count is made until the number of the smallest class is m. Then the remaining $n-m$ consist of the other classes, and if their frequencies are q_1, q_2, q_3, etc., the sum being q, the expected numbers are $\frac{(n-m)q_1}{(n-1)q}$, etc. Hence the following rule may be laid down:

'If a sample is counted until the number of members of one class is m_1, those of the others being then m_2, m_3, etc., and the total n, then the estimate of the frequency of the first class is $\frac{m_1 - 1}{n-1}$, those of the others $\frac{m_2}{n-1}, \frac{m_3}{n-1}$, etc. The variances of these frequencies are approximately $m_r(n-m_r)n^{-3}$.'

A formal proof presents no difficulty.

Similar problems have of course been discussed, and there is a close analogy with problems concerning the duration of play when gambling (cf. Fieller, 1931). Here we have to find the probability that a player starting with a stake of £m, and with a probability p of winning £1 and a probability $1-p$ of losing it per game, will be ruined after n games. This could be generalized to cover the case where the probability of a win was r, of a draw q, and of a loss p; and the problem here considered would be the degenerate case where $r = 0$.

There is also an analogy with one of the methods in use for estimating the frequency of abnormal conditions in human families. These methods have been reviewed by Haldane (1938). If we find m abnormals in a family of n, the expected frequency (on certain assumptions) in a very large family produced by the same parents is not $\frac{m}{n}$ but $\frac{m-1}{n-1}$. For families of this type containing no abnormals would not be recorded, and therefore it is reasonable, as Weinberg (1927) pointed out, to take one of the abnormals as being merely a guarantee that the family is of a type including abnormals, and to base our frequency estimate on his or her sibs.

I have discussed this problem in some detail because I believe that if it is realized that

frequencies can be estimated just as accurately by counting up to a certain number of the rarest type as by counting a certain total, haematologists and others will be saved a good deal of needless effort. I have to thank Drs Kestelmann and Spurway for suggestions.

SUMMARY

If the frequency of an attribute is estimated by counting a sample until m members possessing this attribute are found, and if the total number in the sample is n, then $\frac{m-1}{n-1}$ is an unbiased estimate of the frequency, and its variance is very approximately $\frac{m(n-m)}{n^2(n-1)}$.

REFERENCES

FIELLER, E. C. (1931). The duration of play. *Biometrika*, **22**, 377.

GRÜNEBERG, H. (1942). The anaemia of flexed-tailed mice (*Mus musculus* L.). II. Siderocytes. *J. Genet.* **44**, 246.

HALDANE, J. B. S. (1938). The estimation of the frequencies of recessive conditions in man. *Ann. Eugen., Lond.*, *8*, 255.

WEINBERG, W. (1927). Mathematische Grundlagen der Probanden-methode. *Z. indukt. Abstamm- u. Vererb-Lehre*, **48**, 179.

THE MATHEMATICS OF A POPULATION COMPOSED OF k STATIONARY STRATA EACH RECRUITED FROM THE STRATUM BELOW AND SUPPORTED AT THE LOWEST LEVEL BY A UNIFORM ANNUAL NUMBER OF ENTRANTS

By H. L. SEAL

Consider a population subject to specified stochastic decremental forces operating at each attained age and supported by a uniform annual number of entrants distributed between ages α and β according to a known probability law. It is assumed that the population has existed for at least $\omega-\alpha$ years where ω represents the age, not necessarily finite, at which the expected number of individuals in the population becomes identically zero. The individuals of the, now stationary, population are subdivided into k strata by titular or other distinctions conferred without reference to characteristics influencing the incidence of the decremental forces. The numbers in each of the strata are determined by the action of a stochastic selective force operating on the individuals in stratum g $(g = 1, 2, \ldots, k-1)$, with an intensity which depends only on the length of time the individual has spent in that stratum; if selected the individual moves up into stratum $(g+1)$. It is assumed that the total decremental and selective forces acting at each age and duration, respectively, are invariant with the passage of time.

Write μ_x for the total decremental force applicable stochastically throughout the population at exact age x; this function is supposed continuous and subject to $0<\epsilon<\mu_x$. If ${}_tp_x$ equals the probability that an individual now aged x survives as a member of the population for at least t years then,

$$ {}_tp_x = \exp\left[-\int_0^t \mu_{x+\xi}\,d\xi\right] = \frac{\lambda_{x+t}}{\lambda_x} \quad \text{say} \quad (0 \leqslant t < \omega - x), $$

where $\lambda_x > 0$ may be chosen arbitrarily.

Let $\lambda^{(g)}_{[\xi]}\,d\xi$, $(g = 1, 2, \ldots, k)$ denote the expected number of entrants at any given moment into stratum g at exact age ξ; in particular, $\lambda^{(1)}_{[\xi]}$ is assumed to be a normalized function of bounded variation with $\lambda^{(1)}_{[\xi]} = 0$ $(0 \leqslant \xi < \alpha;\ \xi > \beta)$. Write $p^{(g)}_t$ for the probability that an individual, if alive, remains unselected for at least t years after entry into stratum g, $(g = 1, 2, \ldots, k;\ p^{(k)}_t = 1)$; $p^{(g)}_t$ is thus bounded and monotonically decreasing, and at points of discontinuity it is to be defined by

$$ p^{(g)}_t = \tfrac{1}{2}\{p^{(g)}_{t-0} + p^{(g)}_{t+0}\}. $$

Writing $q^{(g)}_t = 1 - p^{(g)}_t$, $q^{(g)}_t$ is a normalized function of bounded variation in $(0, \omega)$.

It is intended to derive expressions in terms of λ_x, $\lambda^{(1)}_{[x]}$ and $p^{(g)}_t$ $(g = 1, 2, \ldots, k)$, supposed known, for:

(i) $\lambda^{(g)}_{[x]}\,dx$ $(g = 1, 2, \ldots, k)$, the expected number of entrants at any given moment into stratum g at exact age x,

(ii) $A^{(g)}$ $(g = 1, 2, \ldots, k)$, the expected total number of individuals in stratum g at any moment of time.

The expected number of entrants at age x into grade $(g+1)$ is defined to be the aggregate of all selectees for this grade deriving from entrants into grade g at all ages below x; hence, formally,

$$\lambda_{[x]}^{(g+1)} = \int_0^x \lambda_{[x-\xi]\,\xi}^{(g)} p_{x-\xi}\, dq_\xi^{(g)} \quad (g = 1, 2, \ldots, k-1),$$

i.e.
$$\frac{\lambda_{[x]}^{(g+1)}}{\lambda_x} = \int_0^x \frac{\lambda_{[x-\xi]}^{(g)}}{\lambda_{x-\xi}} dq_\xi^{(g)}. \tag{1}$$

The integral on the right-hand side of (1) exists when $g = 1$ and, after defining $\lambda_{[x]}^{(2)}$ suitably at points of discontinuity, $\lambda_{[x]}^{(2)}$ is a normalized function of bounded variation in $(0, \omega)$ (Widder, 1941, Ch. II). By induction $\lambda_{[x]}^{(g)}$ $(g = 3, 4, \ldots, k)$, are normalized functions of bounded variation in $(0, \omega)$.

The expected total of individuals in strata $g+1, g+2, \ldots, k$, is composed of all the survivors from entrants at various ages into grade $(g+1)$; thus

$$\begin{aligned}\sum_{r=g+1}^{k} \Lambda^{(r)} &= \int_0^\omega \lambda_{[x]}^{(g+1)} dx \int_0^{\omega-x} {}_tp_x\, dt \\ &= \int_0^\omega \frac{\lambda_{[x]}^{(g+1)}}{\lambda_x} dx \int_x^\omega \lambda_\xi\, d\xi \\ &= \int_0^\omega \lambda_\xi\, d\xi \int_0^\xi \frac{\lambda_{[x]}^{(g+1)}}{\lambda_x} dx \quad \text{by Dirichlet's integral theorem,}\end{aligned}$$

or
$$\sum_{r=g+1}^{k} \Lambda^{(r)} = \int_0^\omega \lambda_\xi\, d\xi \int_0^\xi (\xi - x)\, d\frac{\lambda_{[x]}^{(g+1)}}{\lambda_x}. \tag{2}$$

Now write, $s = \sigma + i\tau$,

$$f_r(s) = \begin{cases} \displaystyle\int_0^\infty e^{-st}\, dq_t^{(r)} & (r = 1, 2, \ldots, k), \\ \displaystyle\int_0^\infty e^{-st}\, d\frac{\lambda_{[t]}^{(1)}}{\lambda_t} & (r = 0), \end{cases} \tag{3}$$

then (Widder, 1941) if the integrals defining $f_r(s)$ all converge for $\sigma = \sigma_0 > 0$, the rule of multiplication of Laplace transforms results in

$$\int_0^\infty e^{-st}\, d\frac{\lambda_{[t]}^{(g+1)}}{\lambda_t} = \prod_{r=0}^{g} f_r(s) = \phi_g(s) \quad \text{say} \quad (g = 0, 1, 2, \ldots, k-1)$$

and, inverting, $(c > \sigma_0)$,

$$\frac{\lambda_{[x]}^{(g+1)}}{\lambda_x} = \lim_{T\to\infty} \frac{1}{2\pi i} \int_{c-iT}^{c+iT} \frac{e^{sx}}{s} \phi_g(s)\, ds. \tag{4}$$

In a similar manner, if

$$Q_\xi^{(g+1)} = \int_0^\xi (\xi - x)\, d\frac{\lambda_{[x]}^{(g+1)}}{\lambda_x} \quad (g = 0, 1, 2, \ldots, k-1)$$

$$\int_0^\infty e^{-st}\, dQ_t^{(g+1)} = \prod_{r=0}^{g} f_r(s) \int_0^\infty e^{-st}\, dt = \frac{1}{s}\phi_g(s)$$

and $$Q_\xi^{(g+1)} = \lim_{T\to\infty} \frac{1}{2\pi i}\int_{c-iT}^{c+iT} \frac{e^{s\xi}}{s^2}\phi_g(s)\,ds \quad (c>\sigma_0),$$

so that $$\sum_{r=g+1}^{k} \Lambda^{(r)} = \frac{1}{2\pi i}\int_0^\omega \lambda_\xi\, d\xi \int_{c-i\infty}^{c+i\infty} \frac{e^{s\xi}}{s^2}\phi_g(s)\,ds. \tag{5}$$

Example (i). Let $(r = 1, 2, \ldots, k-1)$,

$$-\frac{1}{p_t^{(r)}}\frac{dp_t^{(r)}}{dt} = \begin{cases} c_r & a_r \leqslant t \leqslant b_r, \\ 0 & t < a_r, t > b_r, \end{cases}$$

then $$\frac{dq_t^{(r)}}{dt} = -\frac{dp_t^{(r)}}{dt} = \begin{cases} c_r e^{-c_r(t-a_r)} & a_r \leqslant t \leqslant b_r, \\ 0 & t < a_r,\ t > b_r, \end{cases}$$

and thus $(r = 2, 3, \ldots, k-1)$,

$$f_r(s) = \int_0^\infty e^{-st}\,dq_t^{(r)} = c_r e^{c_r a_r}\int_{a_r}^{b_r} e^{-(s+c_r)t}\,dt = \frac{c_r}{s+c_r}(e^{-sa_r} - R_r e^{-sb_r}),$$

where $R_r = e^{-c_r(b_r-a_r)}$.

It is further assumed that

$$\frac{\lambda_{[t]}^{(1)}}{\lambda_t} = \begin{cases} 1 & t = \alpha, \\ 0 & t \neq \alpha, \end{cases}$$

and since in this case relation (1) does not hold for $g = 1$, $f_0(s)$ must be replaced by unity and $f_1(s)$ given by

$$f_1(s) = \int_0^\infty e^{-st}\,d\frac{\lambda_{[t]}^{(2)}}{\lambda_t},$$

where $$\lambda_{[x]}^{(2)} = \lambda_{[\alpha]}^{(1)}{}_{x-\alpha}p_\alpha \frac{dq_{x-\alpha}^{(1)}}{dx} \quad (\alpha \leqslant x < \omega),$$

i.e. $$\frac{\lambda_{[x]}^{(2)}}{\lambda_x} = \frac{dq_{x-\alpha}^{(1)}}{dx},$$

so that

$$f_1(s) = \int_0^\infty e^{-st}\,d\frac{\lambda_{[t]}^{(2)}}{\lambda_t} = s\int_0^\infty e^{-st}\,dq_{t-\alpha}^{(1)} = se^{-s\alpha}\int_0^\infty e^{-st}\,dq_t^{(1)} = \frac{se^{-s\alpha}c_1}{s+c_1}(e^{-sa_1} - R_1 e^{-sb_1}).$$

Thus $$\phi_g(s) = \frac{C_g s e^{-s\alpha}}{\prod_{r=1}^{g}(s+c_r)}\prod_{r=1}^{g}(e^{-sa_r} - R_r e^{-sb_r}), \quad \text{where } C_g = \prod_{r=1}^{g} c_r$$

$$= \frac{C_g s e^{-s\alpha}}{\prod_{r=1}^{g}(s+c_r)}\sum_{\nu=1}^{2^g} T_\nu e^{-sW_g(\nu)},$$

where $W_g(\nu)$ assumes in succession all the possible sum-combinations of g of the quantities a_r, b_r $(r = 1, 2, \ldots, g)$, without using an a and a b with the same suffix. The coefficient T_ν is equal to the product of those $-R_r$'s corresponding to the b_r's used in the formation of $W_g(\nu)$.

Hence

$$\frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty}\frac{e^{s\xi}}{s^2}\phi_g(s)\,ds = \frac{C_g}{2\pi}\sum_{\nu=1}^{2^g}T_\nu e^{(\xi-\alpha-W_g(\nu))c}\int_{-\infty}^{\infty}\frac{e^{it(\xi-\alpha-W_g(\nu))}}{(c+it)\prod_{r=1}^{g}(c_r+c+it)}\,dt$$

$$= \frac{C_g}{2\pi}\sum_{\nu=1}^{2^g}T_\nu e^{(\xi-\alpha-W_g(\nu))c}\int_{-\infty}^{\infty}e^{it(\xi-\alpha-W_g(\nu))}\sum_{r=0}^{g}\frac{Q_r}{c_r+c+it}\,dt \quad (c_0=0,\ c_r\neq c_j;\ r,j=1,2,\ldots,g)$$

$$= C_g\sum_{\nu=1}^{2^g}T_\nu\sum_{r=0}^{g}Q_r e^{-c_r(\xi-\alpha-W_g(\nu))} \quad (\xi>\alpha+W_g(\nu)),$$

where
$$Q_r = \left\{\prod_{j\neq r}(c_j-c_r)\right\}^{-1} \quad (r=0,1,2,\ldots,g).$$

(A similar type of expression may be found for $\lambda_{[x]}^{(g+1)}/\lambda_{x^g}$.)

And thus, by (5),

$$\sum_{r=g+1}^{k}\Lambda^{(r)} = C_g\int_0^\omega\lambda_x\left\{\sum_{\nu=1}^{2^g}T_\nu\sum_{r=0}^{g}Q_r e^{-c_r(x-\alpha-W_g(\nu))}\right\}dx$$

$$= C_g\sum_{\nu=1}^{2^g}T_\nu\sum_{r=0}^{g}Q_r e^{c_r(\alpha+W_g(\nu))}\overline{N}^{(c_r)}_{\alpha+W_g(\nu)},$$

where
$$\overline{N}^{(c_r)}_{\alpha+W_g(\nu)} = \int_0^{\omega-\alpha-W_g(\nu)}e^{-c_r(\alpha+W_g(\nu)+t)}\lambda_{\alpha+W_g(\nu)+t}\,dt.$$

The above relation holds provided $c_r\neq c_j$ $(r,j=1,2,\ldots,g)$; if, on the other hand, $c_1=c_2=c_3=\ldots=c_g$, it may be shown that

$$\sum_{r=g+1}^{k}\Lambda^{(r)} = \sum_{\nu=1}^{2^g}T_\nu\left\{\overline{N}^{(0)}_{\alpha+W_g(\nu)} - \sum_{r=1}^{g}\frac{c_1^{r-1}}{(r-1)!}\int_0^{\omega-\alpha-W_g(\nu)}t^{r-1}e^{-c_1t}\lambda_{\alpha+W_g(\nu)+t}\,dt\right\}.$$

The case where some of the c's are equal presents no particular difficulties.

The results obtained have a close analogy with those appropriate to a problem in the theory of radioactive transformations (Bateman, 1910).

Example (ii). Let

$$\frac{\lambda_t^{(1)}}{\lambda_t} = \begin{cases}1 & \alpha<t<\beta,\\ \frac{1}{2} & t=\alpha,\ t=\beta,\\ 0 & 0\leqslant t<\alpha,\ t>\beta,\end{cases}$$

and $(r=1,2,\ldots,k-1)$

$$-\frac{dp_t^{(r)}}{dt} = \frac{dq_t^{(r)}}{dt} = \begin{cases}\dfrac{1}{\Gamma(b_r)}e^{-a(t-c_r)}(t-c_r)^{b_r-1} & t>c_r \quad (a,b_r>0),\\ 0 & t<c_r \quad (\alpha+c_1>\beta).\end{cases}$$

Then
$$f_0(s) = e^{-s\alpha}-e^{-s\beta}$$

and
$$f_r(s) = \frac{e^{-sc_r}}{(s+a)^{b_r}} \quad (r=1,2,\ldots,k-1),$$

and thus $(g=0,1,2,\ldots,k-1)$

$$\phi_g(s) = \prod_{r=0}^{g}f_r(s) = \frac{e^{-sC_g^0}-e^{-sC_g^1}}{(s+a)^{B_g}},$$

where

$$C_g^0 = \alpha+\sum_{r=1}^{g}c_r, \quad B_g = \sum_{r=1}^{g}b_r \quad \text{and} \quad C_g^1 = C_g^0-\alpha+\beta \quad (\text{Obs. } C_0^0=\alpha,\ C_0^1=\beta,\ B_0=0).$$

Hence

$$\frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty}\frac{e^{s\xi}}{s^2}\phi_g(s)\,ds=\frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty}\frac{e^{s(\xi-C_g^0)}-e^{s(\xi-C_g^1)}}{s^2(s+a)^{B_g}}\,ds$$

$$=\frac{e^{(\xi-C_g^0)c}}{2\pi}\int_{-\infty}^{\infty}\frac{e^{it(\xi-C_g^0)}}{(c+it)^2(a+c+it)^{B_g}}\,dt-\frac{e^{(\xi-C_g^1)c}}{2\pi}\int_{-\infty}^{\infty}\frac{e^{it(\xi-C_g^1)}}{(c+it)^2(a+c+it)^{B_g}}\,dt.$$

But
$$\frac{e^{xc}}{2\pi}\int_{-\infty}^{\infty}\frac{e^{itx}}{(c+it)^2(a+c+it)^{B_g}}\,dt=\begin{cases}\dfrac{1}{\Gamma(B_g)}\displaystyle\int_0^x e^{-ay}y^{B_g-1}(x-y)\,dy & x>0,\\ 0 & x<0,\end{cases}$$

(see, e.g. § 13·8 of Bochner, 1932) and

$$\int_{C_g^0}^{\omega}\lambda_x\,dx\int_0^{x-C_g^0}e^{-ay}y^{B_g-1}(x-C_g^0-y)\,dy=\int_0^{\omega-C_g^0}\lambda_{\xi+C_g^0}\,d\xi\int_0^{\xi}e^{-ay}y^{B_g-1}(\xi-y)\,dy$$

$$=\int_0^{\omega-C_g^0}e^{-ay}y^{B_g-1}\,dy\int_y^{\omega-C_g^0}(\xi-y)\,\lambda_{\xi+C_g^0}\,d\xi.$$

Writing
$$\bar{S}_z^0=\int_0^{\omega-z}l\lambda_{z+l}\,dl$$

(cp. Steffensen, 1934), there results from (5)

$$\sum_{r=g+1}^{k}A^{(r)}=\frac{1}{\Gamma(B_g)}\int_0^{\omega-C_g^0}e^{-ay}y^{B_g-1}\bar{S}_{y+C_g^0}^0\,dy-\frac{1}{\Gamma(B_g)}\int_0^{\omega-C_g^1}e^{-ay}y^{B_g-1}\bar{S}_{y+C_g^1}^0\,dy.$$

I am very grateful to Dr Stefan Vajda for a number of useful suggestions in connexion with the preceding analysis.

REFERENCES

BATEMAN, H. (1910). The solution of a system of differential equations occurring in the theory of radio-active transformations. *Proc. Camb. Phil. Soc.* **15**, 423.

BOCHNER, S. (1932). *Vorlesungen über Fouriersche Integrale.* Leipzig.

STEFFENSEN, J. F. (1934). *Forsikringsmatematik.* Copenhagen.

WIDDER, D. V. (1941). *The Laplace Transform.* Princeton.

MOMENTS OF r AND χ^2 FOR A FOURFOLD TABLE IN THE ABSENCE OF ASSOCIATION

By J. B. S. HALDANE, F.R.S.

Fourfold contingency tables are in constant use, and it is of interest to calculate the moments of χ^2 derived from them when one or more of the expectations are small, even though the exact method of Fisher (1936) and the table of Fisher & Yates (1938) render this less important than would once have been the case. Further, Kendall's (1942) use of the product-moment correlation of a fourfold table in connexion with rank correlation has given it a new interest. As the third moment of χ^2 and the sixth moment of r are readily derivable as special cases from the formulae of Haldane (1945), it seems worth while to give them, along with the third and fifth moments of r.

Consider the table

a c	b d	L l
M	m	S

$$\chi^2 = \frac{(ad-bc)^2 S}{LlMm}, \quad \text{and} \quad r = \frac{ad-bc}{(LlMm)^{\frac{1}{2}}} = S^{-\frac{1}{2}}\chi.$$

If L and l or M and m are samples of the same population falling into two classes, then $\bar{r} = 0$, and $\overline{\chi^2} = 1$ approximately. The exact values of the first three moments of χ^2 are readily derived from equations (1) of Haldane (1945), putting

$$n = 2, \quad k = \frac{S^2}{Ll}, \quad R_1 = \frac{S}{Mm}, \quad R_2 = \frac{S^2}{M^2m^2} - \frac{2}{Mm}.$$

Let $LlMm = \lambda^2$, $Ll + Mm = \mu$, $(L-l)(M-m) = \nu$. Then

$$\overline{\chi^2} = \frac{S}{S-1},$$

$$\overline{\chi^4} = \frac{S^2[3(S+6)\lambda^2 - 6S^2\mu + S^3(S+1)]}{\lambda^2(S-1)(S-2)(S-3)},$$

$$\overline{\chi^6} = \frac{S^3}{\lambda^4(S-1)(S-2)(S-3)(S-4)(S-5)}[5(3S^2+86S+120)\lambda^4 - 10S^2(13S+60)\lambda^2\mu + 120S^4\mu^2 - 5S^2(5S^2+87S+60)\lambda^2 - 30S^5(S+3)\mu + S^5(S+1)(S^2+15S-4)],$$

$$\kappa_1 = \mu_1' = \frac{S}{S-1},$$

$$\kappa_2 = \mu_2 = \frac{S^2}{\lambda^2(S-1)(S-2)(S-3)}[2(S-1)^{-1}(S^2+10S-12)\lambda^2 - 6S^2\mu + S^3(S+1)],$$

$$\kappa_3 = \mu_3 = \frac{S^3}{\lambda^4(S-1)(S-2)(S-3)(S-4)(S-5)}[8(S-1)^{-2}(S^4+51S^3+22S^2-308S+240)\lambda^4 - 8(S-1)^{-1}S^2(14S^2+79S-120)\lambda^2\mu + 120S^4\mu^2 - 2(S-1)^{-1}S^3(14S^3+193S^2-51S-120)\lambda^2 - 30S^5(S+3)\mu + S^5(S+1)(S^2+15S-4)]. \quad (1)$$

The odd moments of r can be calculated as follows. To obtain the mean value of an odd power of $(ad-bc)$ we use the operator Δ. This permutes the indices of a and d, and also those of b and c, without change of sign. It also permutes the indices (normal and factorial) of a and d with those of b and c, the sign being changed. No term is repeated if obtained more than once, and all are added together. Thus:

$$\Delta(a^3d^2) = a^3d^2 + a^2d^3 - b^3c^2 - b^2c^3,$$

$$\Delta(a^3bd^2) = a^3bd^2 + a^2bd^3 + a^3cd^2 + a^2cd^3 - ab^3c^2 - b^3c^2d - ab^2c^3 - b^2c^3d.$$

Now the mean value of a product of factorials of a, b, c, d is

$$E[a^{(\alpha)}b^{(\beta)}c^{(\gamma)}d^{(\delta)}] = E[a^{(\alpha+\epsilon)}b^{(\beta-\epsilon)}c^{(\gamma-\epsilon)}d^{(\delta-\epsilon)}]$$
$$= L^{(\alpha+\beta)}l^{(\gamma+\delta)}M^{(\alpha+\gamma)}m^{(\beta+\delta)}S^{-(\alpha+\beta+\gamma+\delta)},$$

where $a^{(\alpha)} \equiv a(a-1)(a-2)\ldots(a-\alpha+1)$, and so on (Haldane, 1940). It follows that

$$E\Delta[a^{(\alpha)}b^{(\beta)}c^{(\gamma)}d^{(\delta)}] = 0, \quad \text{if} \quad \alpha+\beta = \gamma+\delta \text{ or } \alpha+\gamma = \beta+\delta,$$

and
$$E\Delta[a^{(\alpha)}d^{(\delta)}] = S^{-(\alpha+\delta)}[L^{(\alpha)}l^{(\delta)} - L^{(\delta)}l^{(\alpha)}][M^{(\alpha)}m^{(\delta)} - M^{(\delta)}m^{(\alpha)}],$$
$$E\Delta[a^{(\alpha)}bd^{(\delta)}] = S^{-(\alpha+\delta+1)}[\{L^{(\alpha+1)}l^{(\delta)} - L^{(\delta)}l^{(\alpha+1)}\}\{M^{(\alpha)}m^{(\delta+1)} - M^{(\delta+1)}m^{(\alpha)}\}$$
$$+ \{L^{(\alpha)}l^{(\delta+1)} - L^{(\delta+1)}l^{(\alpha)}\}\{M^{(\alpha+1)}m^{(\delta)} - M^{(\delta)}m^{\alpha+1}\}],$$
$$E\Delta[a^{(\alpha)}bd^{(\alpha)}] = -E\Delta[a^{(\alpha+1)}d^{(\alpha)}].$$

Thus $(ad-bc)^3 = \Delta(a^3d^3 - 3a^2bcd^2)$
$$= \Delta[a^{(3)}d^{(3)} - 3a^{(2)}bcd^{(2)} + 3a^{(3)}d^{(2)} - 3a^{(2)}bcd + a^{(3)}d + 9a^{(2)}d^{(2)} + 3a^{(2)}d + ad].$$

So $E[(ad-bc)^3] = E\Delta[a^{(3)}d + 3a^{(2)}d]$, the other terms vanishing,
$$= S^{-(4)}[L^{(3)}l - Ll^{(3)}][M^{(3)}m - Mm^{(3)}] + 3S^{-(3)}[L^{(2)}l - Ll^{(2)}][M^{(2)}m - Mm^{(2)}]$$
$$= (S-1)^{-(2)}\lambda^2\nu.$$

In general, a^x must be expanded in factorials as a polynomial whose coefficients are the initial differences of the powers of integers, divided by the appropriate factorials.

$$(ad-bc)^5 = \Delta(a^5d^5 - 5a^4bcd^4 + 10a^3b^2c^2d^3)$$
$$= \Delta[a^{(5)}d^{(5)} - 5a^{(4)}bcd^{(4)} + 10a^{(3)}b^{(2)}c^{(2)}d^{(3)}$$
$$+ 10\{a^{(5)}d^{(4)} - 3a^{(4)}bcd^{(3)} + 3a^{(3)}b^{(2)}c^{(2)}d^{(2)} + a^{(3)}b^{(2)}cd^{(3)}\}$$
$$+ 5\{5a^{(5)}d^{(3)} + 20a^{(4)}d^{(4)} - 7a^{(4)}bcd^{(2)} - 34a^{(3)}bcd^{(3)} + 2a^{(3)}b^{(2)}c^{(2)}d + 6a^{(3)}b^{(2)}cd^{(2)}$$
$$+ 18a^{(3)}b^{(2)}c^{(2)}d^{(2)}\}$$
$$+ 5\{3a^{(5)}d^{(2)} + 50a^{(4)}d^{(3)} - a^{(4)}bcd^{(2)} - 36a^{(3)}bcd^{(2)} - 12a^{(2)}b^{(2)}c^{(2)}d + 2a^{(3)}b^{(2)}cd\}$$
$$+ \{a^{(5)}d + 150a^{(4)}d^{(2)} + 625a^{(3)}d^{(3)} - 20a^{(3)}bcd - 165a^{(2)}bcd^{(2)} + 30a^{(2)}b^{(2)}cd\}$$
$$+ 5\{2a^{(4)}d + 75a^{(3)}d^{(2)} - 3a^{(2)}bcd\} + 25\{a^{(3)}d + 9a^{(2)}d^{(2)}\} + 15a^{(2)}d].$$

So $E[(ad-bc)^5] = E\Delta[10\{a^{(5)}d^{(2)} + a^{(4)}d^{(3)} + a^{(4)}bd^{(2)}\} + \{a^{(5)}d + 130a^{(4)}d^{(2)}\}$
$$+ 10\{a^{(4)}d + 36a^{(3)}d^{(2)}\} + 25a^{(3)}d + 15a^{(2)}d]$$
$$= 10S^{-(7)}[\{L^{(5)}l^{(2)} - L^{(2)}l^{(5)}\}\{M^{(5)}m^{(2)} - M^{(2)}m^{(5)}\} + \{L^{(4)}l^{(3)} - L^{(3)}l^{(4)}\}\{M^{(4)}m^{(3)} - M^{(3)}m^{(4)}\}$$
$$+ \{L^{(3)}l^{(2)} - L^{(2)}l^{(5)}\}\{M^{(4)}m^{(3)} - M^{(3)}m^{4}\} + \{L^{(4)}l^{(3)} - L^{(3)}l^{(4)}\}\{M^{(5)}m^{(2)} - M^{(2)}m^{(5)}\}]$$
$$+ S^{-(6)}[\{L^{(5)}l - Ll^{(5)}\}\{M^{(5)}m - Mm^{(5)}\} + 130\{L^{(4)}l^{(2)} - L^{(3)}l^{(4)}\}\{M^{(4)}m^{(2)} - M^{(2)}m^{(4)}\}]$$
$$+ 10S^{-(5)}[\{L^{(4)}l - Ll^{(4)}\}\{M^{(4)}m - Mm^{(4)}\} + 36\{L^{(3)}l^{(2)} - L^{(2)}l^{(3)}\}\{M^{(3)}m^{(2)} - M^{(2)}m^{(3)}\}]$$
$$+ 25S^{-(4)}[L^{(3)}l - Ll^{(3)}][M^{(3)}m - Mm^{(3)}] + 15[L^{(2)}l - Ll^{(2)}][M^{(2)}m - Mm^{(2)}]$$

$$= 10S^{-(7)}\lambda^2\nu[\lambda^2-(S-1)\mu+(S-1)^2](S-5)^2(S-6)^2$$
$$+ S^{-(6)}\lambda^2\nu[134\lambda^2-2(S^2+60S-55)\mu+S^4-10S^3+175S^2-360S+230](S-5)^2$$
$$+ 10S^{-(5)}\lambda^2\nu[37\lambda^2-(S^2+30S-25)\mu+S^4-12S^3+94S^2-204S+157]$$
$$+ 25S^{-(4)}\lambda^2\nu(S-3)^2+15S^{-(3)}\lambda^2\nu$$
$$= (S-1)^{-(6)}\lambda^2\nu[2(5S+12)\lambda^2-12S^2\mu+S^3(S+5)].$$

Hence the moments of r are

$$\mu_1' = 0,$$

$$\mu_2 = \frac{1}{S-1},$$

$$\mu_3 = \frac{\nu}{(S-1)(S-2)},$$

$$\mu_4 = \frac{3(S+6)\lambda^2-6S^2\mu+S^3(S+1)}{\lambda^2(S-1)(S-2)(S-3)},$$

$$\mu_5 = \frac{\nu[2(5S+12)\lambda^2-12S^2\mu+S^3(S+5)]}{\lambda^3(S-1)(S-2)(S-3)(S-4)},$$

$$\mu_6 = [5(3S^2+86S+120)\lambda^4-10S^2(13S+60)\lambda^2\mu+120S^4\mu^2-5S^2(5S^2+87S+60)\lambda^3$$
$$-30S^5(S+3)\mu+S^5(S+1)(S^2+15S-4)]\div\lambda^4(S-1)(S-2)(S-3)(S-4)(S-5),$$

$$\kappa_4 = \frac{\dfrac{6(5S-6)\lambda^2}{S-1}-6S^2\mu+S^3(S+1)}{\lambda^2(S-1)(S-2)(S-3)(S-4)},$$

$$\kappa_5 = \frac{\nu\left[\dfrac{12(7S-12)\lambda^2}{S-1}-12S^2\mu+S^3(S+5)\right]}{\lambda^3(S-1)(S-2)(S-3)(S-4)}.$$

The odd moments vanish if $L = l$, or $M = m$, i.e. if the samples are equal, or the class frequencies 50%. κ_4 is negative if L and l, or M and m, are nearly equal, but becomes positive if both a class frequency and a sample are small.

These expressions may be used for accurate tests of the significance of observed values of χ^2 or r, or, which is more difficult by the methods mentioned above, the significance of a series of values of these parameters.

REFERENCES

FISHER, R. A. (1936). The Logic of inductive inference. *J. Roy. Statist. Soc.* **98**, 39–82.

FISHER, R. A. & YATES, F. (1938). *Statistical Tables for Biological, Medical and Agricultural Research.* Edinburgh.

HALDANE, J. B. S. (1937). The exact value of the moments of the distribution of χ^2, used as a test of goodness of fit when expectations are small. *Biometrika*, **29**, 389–91.

HALDANE, J. B. S. (1940). The mean and variance of χ^2, when used as a test of homogeneity, when expectations are small. *Biometrika*, **31**, 346–55.

HALDANE, J. B. S. (1945). The use of χ^2 as a test of homogeneity in a $(n \times 2)$-fold table when expectations are small. *Biometrika*, **33**, 234.

KENDALL, M. G. (1942). Partial rank correlation. *Biometrika*, **32**, 277–83.

THE USE OF χ^2 AS A TEST OF HOMOGENEITY IN A $(n\times 2)$-FOLD TABLE WHEN EXPECTATIONS ARE SMALL

By J. B. S. HALDANE, F.R.S.

The χ^2 test has proved so useful that any extension of its field of application is likely to be of value. Its use is at present deprecated when the expectation in any group is small. This may occur either because samples are small, or because one of the classes expected in each sample is rare. Fisher (1941) recommends 5 as a lower limit for expectations. It will be shown that the test may still be applied even when the expectation is less than unity. Where the number of classes is large, we may still sometimes use the ordinary tables. Otherwise the value of P must be calculated in each particular case. Haldane (1937) showed how this could be done when χ^2 is used as a test of goodness of fit. But, particularly in genetical work, it has found its greatest use as a test of homogeneity. The formulae required for a $(m \times n)$-fold table are very cumbrous. Those for a $(n \times 2)$-fold table are developed below. The value of the variance has already been given (Haldane, 1940). That of the third moment is now calculated.

Consider a set of n samples of $s_1, s_2, \ldots, s_i, \ldots, s_n$ individuals. In the ith sample let there be a_i individuals of class X and b_i of class Y. Let $\Sigma a_i = A$, $\Sigma b_i = B$, $\Sigma s_i = S$, so that the table is

a_1	a_2	...	a_i	...	a_n	A
b_1	b_2	...	b_i	...	b_n	B
s_1	s_2	...	s_i	...	s_n	S

Let $\qquad \Sigma s_i^{-1} = R_1, \quad \Sigma s_i^{-2} = R_2, \quad S^2/AB = k.$

Then it can readily be shown that $\chi^2 = S\left(1 - \dfrac{Sx}{AB}\right)$, where $x = \Sigma\left(\dfrac{a_i b_i}{s_i}\right)$. It also follows at once from the argument of Haldane (1940, p. 347) that in a homogeneous population the expectation of $a_i^{(\alpha)} b_i^{(\beta)}$ is $S^{-(\alpha+\beta)} A^{(\alpha)} B^{(\beta)} s_i^{(\alpha+\beta)}$, where $y^{(\alpha)} = y(y-1)(y-2)\ldots(y-\alpha+1)$. Hence

$$\bar{x} = S^{-(1)} AB(S-n),$$

$$\overline{x^2} = S^{-(4)} A^2 B^2 [S^2 - 2(n+2)S + n(n+10) - 6R_1 + kS^{-2} \times \{-(n^2+2n-2)S + n(n-2) + R_1 S(S+1)\}].$$

The proof is similar to that which follows for $\overline{x^3}$, in which multiple summations are taken over all unequal sets of values of i, j, and i, j, k.

$$\begin{aligned} x^3 &= \Sigma s_i^{-3} a_i^3 b_i^3 + 3\Sigma\Sigma s_i^{-2} a_i^2 b_i^2 s_j^{-1} a_j b_j + 6\Sigma\Sigma\Sigma s_i^{-1} a_i b_i s_j^{-1} a_j b_j s_k^{-1} a_k b_k \\ &= \Sigma s_i^{-3} [a_i^{(3)} b_i^{(3)} + 3a_i^{(2)} b_i^{(2)} + 3a_i^{(2)} b_i^{(2)} + a_i^{(3)} b_i + 9a_i^{(2)} b_i^{(2)} + a_i b_i^{(3)} + 3a_i^{(2)} b_i + 3a_i b_i^{(2)} + a_i b_i] \\ &\quad + 3\Sigma\Sigma s_i^{-2} s_j^{-1} [a_i^{(2)} b_i^{(2)} a_j b_j + a_i^{(2)} b_i a_j b_j + a_i b_i^{(2)} a_j b_j + a_i b_i a_j b_j] \\ &\quad + 6\Sigma\Sigma\Sigma s_i^{-1} a_i b_i s_j^{-1} a_j b_j s_k^{-1} a_k b_k. \end{aligned}$$

$$\begin{aligned}
\text{Hence } \overline{x^3} &= S^{-(6)}A^{(3)}B^{(3)}[\Sigma s_i^{-3}s_i^{(6)} + 3\,\Sigma\Sigma s_i^{-2}s_i^{(4)}s_j^{-1}s_j^{(2)} + 6\Sigma\Sigma\Sigma s_i^{-1}s_i^{(2)}s_j^{-1}s_j^{(2)}s_k^{-1}s_k^{(2)}] \\
&\quad + 3S^{-(5)}[A^{(3)}B^{(2)} + A^{(2)}B^{(3)}][\Sigma s_i^{-3}s_i^{(5)} + \Sigma\Sigma s_i^{-2}s_i^{(2)}s_j^{-1}s_j^{(3)}] \\
&\quad + S^{-(4)}[A^{(3)}B + 9A^{(2)}B^{(2)} + AB^{(3)}]\,\Sigma s_i^{-3}s_i^{(4)} + 3S^{-(4)}A^{(2)}B^{(2)}\,\Sigma\Sigma s_i^{-2}s_i^{(2)}s_j^{-2}s_j^{(2)} \\
&\quad + 3S^{-(3)}[A^{(2)}B + AB^{(2)}]\,\Sigma s_i^{-3}s_i^{(3)} + s_i^{-(2)}AB\,\Sigma s_i^{-3}s_i^{(2)} \\
&= S^{-(6)}AB(AB - S + 1)(AB - 2S + 4)[\Sigma(s_i^3 - 15s_i^2 + 85s_i - 225 + 274s_i^{-1} - 120s_i^{-2}) \\
&\quad + 3\Sigma\Sigma(s_i^2 - 6s_i + 11 - 6s_i^{-1})(s_j - 1) + 6\Sigma\Sigma\Sigma(s_i - 1)(s_j - 1)(s_k - 1)] \\
&\quad + 3S^{-(5)}AB(AB - S + 1)(S - 4)[\Sigma(s_i^2 - 10s_i + 35 - 50s_i^{-1} \\
&\quad + 24s_i^{-2}) + \Sigma\Sigma(s_i - 3 + 2s_i^{-1})(s_j - 1)] \\
&\quad + S^{-(4)}AB(7AB + S^2 - 12S + 13)\,\Sigma(s_i - 6 + 11s_i^{-1} - 6s_i^{-2}) \\
&\quad + 3S^{-(4)}AB(AB - S + 1)\,\Sigma\Sigma(1 - s_i^{-1})(s_j - 1) \\
&\quad + 3S^{-(3)}AB(S - 2)\,\Sigma(1 - 3s_i^{-1} + 2s_i^{-2}) + S^{-(2)}AB\Sigma(s_i^{-1} - s_i^{-2}).
\end{aligned}$$

$$\begin{aligned}
\text{But } \quad & \Sigma(s_i^3 - 3s_i^2 + 3s_i - 1) + 3\Sigma\Sigma(s_i^2 - 2s_i + 1)(s_j - 1) + 6\Sigma\Sigma\Sigma(s_i - 1)(s_j - 1)(s_k - 1) \\
&= [\Sigma(s_i - 1)]^3 = (S - n)^3, \\
& \Sigma(s_i^2 - 2s_i + 1) + 2\Sigma\Sigma(s_i - 1)(s_j - 1) = [\Sigma(s_i - 1)]^2 = (S - n)^2, \\
& \Sigma(s_i - 2 + s_i^{-1}) + 2\Sigma\Sigma(1 - s_i^{-1})(s_j - 1) = \Sigma(s_i - 1)\Sigma(1 - s_i^{-1}) = (S - n)(n - R_1).
\end{aligned}$$

$$\begin{aligned}
\text{So } \quad \overline{x^3} &= S^{-(6)}AB[A^2B^2 - (3S - 5)\,AB + 2(S - 1)(S - 2)] \\
&\quad \times [(S - n)^3 - 12(S - n)^2 + 18(S - n)(n - R_1) + 40S - 176n + 256R_1 - 120R_2] \\
&\quad + 3S^{-(4)}AB(AB - S + 1)[(S - n)^2 - (S - n)(n - R_1) - 7S + 32n - 49R_1 + 24R_2] \\
&\quad + S^{-(4)}AB(7AB + S^2 - 12S + 13)(S - 6n + 11R_1 - 6R_2) + S^{-(2)}AB(3n - 8R_1 + 5R_2) \\
&= S^{-4}S^{-(6)}A^3B^3[S^4\{S^3 - 3(n + 4)\,S^2 + (3n^2 + 42n + 40)\,S - n(n^3 + 30n + 176)\} \\
&\quad - kS^2\{3(n^2 + 2n - 2)\,S^2 - (3n^3 + 51n^2 + 72n - 80)\,S + 5n(n^2 + 6n - 40)\} \\
&\quad - 2k^2\{(n^3 + 9n^2 + 14n - 12)\,S^2 - 3(n - 2)(n + 5)\,S - 2n(n^2 - 4)\} \\
&\quad + 32R_1\{6S^4 - kS^2(S - 1)(S + 10) + k^2(S - 1)^2(S + 4)\} \\
&\quad - R_1\{2S^4(9S - 128) - kS^2(3S^3 - 43S^2 - 168S - 120) - 2k^2(11S^3 + 23S^2 + 10S - 4)\} \\
&\quad - R_2S\{120S^3 - 30kS^2(S + 3) + k^2(S + 1)(S^2 + 15S - 4)\}].
\end{aligned}$$

$$\text{But } \qquad \overline{\chi^4} = S^2\left(1 - \frac{2S\overline{x}}{AB} - \frac{S^2\overline{x^2}}{A^2B^2}\right), \quad \overline{\chi^6} = S^3\left(1 - \frac{3S\overline{x}}{AB} + \frac{3S^2\overline{x^2}}{A^2B^2} - \frac{S^3\overline{x^3}}{A^3B^3}\right).$$

$$\begin{aligned}
\text{Hence } \overline{\chi^2} &= S(S - 1)^{-1}(n - 1), \\
\overline{\chi^4} &= S(S - 1)^{-(3)}[(n^2 - 1)\,S^2 + 6(2n - 1)\,S - 6R_1S^2 - k\{-(n^2 + 2n - 2)\,S \\
&\quad + n(n - 2) + R_1S(S + 1)\}], \\
\overline{\chi^6} &= S(S - 1)^{-(5)}[S^2\{(n - 1)(n + 1)(n + 3)\,S^2 + 2(30n^2 + 69n - 43)\,S + 120(3n - 1)\} \\
&\quad - kS\{(3n^3 + 21n^2 + 24n - 26)\,S^2 - (5n^3 - 57n^2 - 266n + 120)\,S - 60n(n - 2)\} \\
&\quad + 2k^2\{(n^3 + 9n^2 + 14n - 12)\,S^2 - 3n(n - 2)(n + 5)\,S + 2n(n^2 - 4)\} \\
&\quad - 3nR_1\{6S^4 - kS^2(S - 1)(S + 10) + k^2(S - 1)^2(S + 4)\} \\
&\quad - R_1\{2S^3(47S + 180) - kS^2(19S^2 + 201S + 180) + 2k^2(11S^3 + 23S^2 + 10S - 8)\} \\
&\quad + R_2S\{120S^3 - 30kS^2(S + 3) + k^2(S + 1)(S^2 + 15S - 4)\}]. \qquad (1)
\end{aligned}$$

These equations are rarely the most convenient. It is usually better to subtract the classical values, and write

$$\begin{aligned}
\overline{\chi^2} &= n-1+(S-1)^{-1}(n-1),\\
\overline{\chi^4} &= n^2-1+(S-1)^{-(3)}[(k-6)\,S\{R_1S(S+1)-(n^2+2n-2)\,S+n(n-2)\}\\
&\quad +6R_1S^2-(5n^2+12n-11)\,S+6(n^2-1)],\\
\overline{\chi^6} &= (n-1)(n+1)(n+3)+(S-1)^{-(5)}[S^3\{(5-k)(3n^3+21n^2+24n-26)\,S\\
&\quad +(3n-1)(S+120)\}-(n-1)(n+1)(n+3)(85S^3-225S^2+274S-120)\\
&\quad +kS^2\{(5n^3-57n^2-266n+120)\,S-60n(n-2)\}\\
&\quad +2k^2S\{(n^3+9n^2+14n-12)\,S^2-3n(n-2)(n+5)\,S+2n(n^2-4)\}\\
&\quad -3nR_1S\{6S^4-kS^2(S-1)(S+10)+k^2(S-1)^2(S+4)\}\\
&\quad -R_1S\{2S^3(47S+180)-kS^2(19S^2+210S+180)+2k^2(11S^3+23S^2+10S-8)\}\\
&\quad +R_2S^3\{120S^3-30kS^2(S+3)+k^2(S+1)(S^2+15S-4)\}]. \qquad (2)
\end{aligned}$$

To calculate the moments about the mean, we write

$$\overline{\chi^2}=n-1+\alpha,\quad \overline{\chi^4}=n^2-1+\beta,\quad \overline{\chi^6}=(n-1)(n+1)(n+3)+\gamma,$$

whence $\kappa_2=\mu_2=2(n-1)-2(n-1)\alpha+\beta-\alpha^2$,

$$\kappa_3=\mu_3=8(n-1)+3(n-1)(n-3)\alpha-3(n-1)(\beta-2\alpha^2)+\gamma-3\alpha\beta+2\alpha^3. \qquad (3)$$

The full algebraical expressions for these moments are rather cumbrous. However, by expanding equations (2) in descending powers of S, we have

$$\begin{aligned}
\kappa_1 &= \mu_1'=n-1+(n-1)\,S^{-1}+\ldots,\\
\kappa_2 &= \mu_2=2(n-1)+(k-6)\,R_1-[k(n^2+2n-2)-2(2n^2+8n-7)]\,S^{-1}+\ldots,\\
\kappa_3 &= \mu_3=8(n-1)+2(11k-56)\,R_1+(k^2-30k+120)\,R_2\\
&\quad -2(3n-2)\,[(3n+8)\,k-4(3n+11)]\,S^{-1}+\ldots. \qquad (4)
\end{aligned}$$

In these equations the comparatively small terms involving R_1 and R_2 are omitted in the coefficients of S. The easiest forms for computation are

$$\begin{aligned}
\kappa_2 &= 2(n-1)+(k-6)\,R_1-[(k-4)(n^2+2n-2)-2(4n+1)]\,S^{-1},\\
\kappa_3 &= 8(n-1)+2[11(k-5)-1]\,R_1+[(k-15)^2-105]\,R_2\\
&\quad -2(3n-2)\,[(k-4)(3n+8)-12]\,S^{-1}.
\end{aligned}$$

These equations are sufficiently accurate for most practical applications.

The terms in equations (4) not involving negative powers of S are identical with the moments of χ^2 for $n-1$ degrees of freedom, used as a test of goodness of fit, derivable from Haldane's (1937) equations (3), if the expectations of a_r and b_r are ps_r and qs_r, and $pq=k^{-1}$. The effect of using χ^2 as a test of homogeneity rather than goodness of fit is to reduce the variance and the positive skewness, the reduction being considerable if k is large. There can also be little doubt, by analogy with the simpler case, that the fourth cumulant of χ^2 as a test of homogeneity is

$$\begin{aligned}
\kappa_4 &= 48(n-1)+96(4k-19)\,R_1+16(7k^2-125k+420)\,R_2\\
&\quad +(k^3-126k^2+1680k-5040)\,R_3+O(S^{-1}).
\end{aligned}$$

It will be seen that when $k-4$ is small, that is to say, the two types are almost equally frequent, both variance and skewness are reduced if samples are small. But when one class is fairly rare, so that k is large, they may be considerably increased.

In any particular case, one of three things may happen. It may be clear from equations (4) that the value of χ^2 is not significantly above that expected on a basis of homogeneity. This will certainly be so if $\chi^2-(n-1)$ is less than $\kappa_2^{\frac{1}{2}}$, and in this case there is no need to calculate κ_3. It may be clear that the corrections to the moments are unimportant, in which case the ordinary tables may be used. This is especially frequent for large values of n, where the distribution of χ^2 tends to normality. Or finally it is necessary to take κ_3 into consideration. If so we calculate κ_2 and κ_3 from equations (2), and make the transformation (Haldane, 1938)

$$\xi=\left[\left(\frac{\chi^2}{\kappa_1}\right)^h+\frac{h(1-h)\kappa_1}{2\kappa_1^2}-1\right]\frac{\kappa_1}{h\kappa_2^{\frac{1}{2}}}, \tag{5}$$

where $h=1-\frac{\kappa_1\kappa_3}{3\kappa_2^2}$. Then ξ is almost normally distributed with mean zero and variance unity. The probability that χ^2 should exceed its observed value is the probability that ξ should do so.

MOMENTS OF χ^2 WHEN ALL SAMPLES ARE EQUAL

If every s_r is equal, then $SR_1=n^2$, $S^2R_2=n^3$. So equations (1) become

$$\begin{aligned}
\overline{\chi^2}&=(n-1)\,S(S-1)^{-1},\\
\overline{\chi^4}&=(n-1)\,S(S-1)^{-(3)}[(n-1)\,S^2-6(n-1)\,S-2k(S-n)],\\
\overline{\chi^6}&=(n-1)\,(S-1)^{-(5)}[S^3\{(n+1)\,(n+3)\,S^2-2(9n^2+26n-43)\,S+120(n-1)^2\}\\
&\quad-2kS^2(S-n)\{(n+13)\,S-60(n-1)\}-4k^2(S-n)\{(n-6)\,S^2+9nS-4n\}],\\
\mu_2&=2(n-1)\,S(S-n)\,(S-1)^{-(3)}\left(\frac{S^2}{S-1}-k\right),\\
\mu_3&=2(n-1)\,(S-n)\,(S-1)^{-(5)}[4S^4(S-1)^{-2}\{S^2-(4n-5)\,S-2n\}\\
&\quad-kS^3(S-1)^{-1}\{(n-8)\,S-17n+10\}-4k^2\{(n-6)\,S^2+9nS-4n\}].
\end{aligned} \tag{6}$$

These equations, and those of the following paper, are a useful check on the accuracy of equations (1). They show that when the numbers in the samples are equal, or nearly so, the rarity of one class will always reduce μ_2, and will reduce μ_3 unless n is less than 8. However, μ_3 is always positive.

A NUMERICAL APPLICATION

Spurway (1945) investigated the frequency of crossing-over between a number of sex-linked genes in *Drosophila subobscura*, and used χ^2 as a test of the homogeneity of different families. In most cases no expectation was less than 5, and the classical method was used. However, for 9 pairs of genes the expectations were often very small, usually because crossing-over was infrequent. Her results are tabulated in Table 1. In the first column *C* denotes that the genes were in coupling, *R* that they were in repulsion, n is the number of families, S the total flies in them, A the number of flies showing crossing-over. Table 2 shows the actual data for the genes *bg* and *ct*, tabulated in the last row.

κ_2 was calculated from formulae (2) and (3), so the values are exact. It would have been quite sufficient to use formula (4). For example, in the case of *wi* and *y* this gives 234·9 instead of 235·8, and in that of *bg ct*, 17·570 in place of 17·507. The values of κ_2 are considerably greater than the classical value $2(n-1)$ in some cases, slightly less in others. κ_3 is calculated from formula (4) throughout. It may also be increased up to about 12 times, or slightly decreased. The accurate formula is not required, since, when n is large, a comparatively large change in κ_3 does not greatly affect P.

The values of P are calculated from equation (5) except when P is nearly 0·5, when the distribution is taken as normal. It will be seen that none of the values of χ^2 are significantly larger than their expectation, nor is their total. If no correction had been made, but the

Table 1. *Data on crossing-over in* Drosophila subobscura. *Explanation in text*

Genes	n	χ^2	S	A	R_1	R_2	κ_1	κ_2	κ_3	P
wi y C	63	83·245	1342	18	4·7266	0·56584	62·047	235·838	5915	0·10
cv l C	25	24·563	776	35	1·4129	0·12783	24·031	56·034	454·7	0·47
cv l R	66	64·662	1656	88	4·2961	0·49543	65·040	149·245	1106	0·51
l sn C	39	47·259	1167	386	2·2898	0·28976	38·033	72·162	266·1	0·14
l sn R	22	20·410	694	283	1·2712	0·12315	21·030	39·764	143·3	0·54
cp sn C	40	39·959	2359	113	1·1336	0·06469	39·017	83·467	499·8	0·46
cp sn R	103	121·970	5617	300	2·9840	0·17773	102·018	214·916	1220	0·09
sn bg C	10	6·523	273	45	0·8752	0·16451	9·033	17·993	83·43	0·25
bg ct C	10	9·231	273	54	0·8752	0·16451	9·033	17·507	74·85	0·48
	378	417·82					369·28	886·87	9763	0·07

Table 2. *Numbers of male flies, s_r, and cross-overs, a_r, between the loci of bg and ct, in ten families.* $\Sigma s_r^{-1} = 0{\cdot}8752$, $\Sigma s_r^{-2} = 0{\cdot}1645$. *Smallest expectation of a_r is* 0·593

s_r	23	68	28	12	28	7	7	38	3	59	273
a_r	2	18	3	2	7	0	1	10	1	10	54

usual distribution of χ^2 had been employed, the values of P for *wi y*, *cp sn R*, and the total, would have been 0·035, 0·06 and 0·04 respectively. Thus the data would have been judged significantly heterogeneous.

Summary

χ^2 can be used as a test of homogeneity, even when expectations are less than unity, by the use of the formulae here given. In many cases the approximate formulae (4) are quite sufficient.

I have to thank Dr Spurway for the use of her numerical data, and for help with the computations.

REFERENCES

Fisher, R. A. (1941). *Statistical Methods for Research Workers.* Edinburgh.

Haldane, J. B. S. (1937). The exact value of the moments of the distribution of χ^2, used as a test of goodness of fit when expectations are small. *Biometrika*, **29**, 389–91.

Haldane, J. B. S. (1938). The approximate normalization of a class of frequency distributions. *Biometrika*, **29**, 392–404.

Haldane, J. B. S. (1940). The mean and variance of χ^2, when used as a test of homogeneity, when expectations are small. *Biometrika*, **31**, 346–55.

Spurway, Helen (1945). The genetics and cytology of *Drosophila subobscura*. 1. Element A. Sex-linked mutants and their standard order. *J. Genet.* **46**, 268–86.

THE TREATMENT OF TIES IN RANKING PROBLEMS

By M. G. KENDALL

1. When a number of objects are presented for ranking by an observer there sometimes arise cases in which he is unable to express a preference in regard to certain of them and 'ranks them equal' or regards them as 'tying'. The effect may arise either because the objects really are indistinguishable, so far as the quality under consideration is concerned, or because the observer is unable to discern such differences as exist. Ties of this character are by no means uncommon—and indeed may be more the rule than the exception in some classes of work—and it is desirable to have a systematic method of dealing with them. In this paper I consider the effect of ties on coefficients of rank correlation, the estimation of rankings and the measurement of concordance in judges.

Rank correlations

2. The method of allocating ranking numbers to tied individuals in general use is to average the ranks which they cover. For instance, if the observer ties the third and fourth members each is allotted the number $3\frac{1}{2}$, and if he ties the second to the seventh inclusive, each is allotted the number $\frac{1}{6}(2+3+4+5+6+7) = 4\frac{1}{2}$. This is known as the mid-rank method and is the only one I shall consider. In fact I have seen only two other courses mentioned:

(*a*) 'Student' (1921) refers to a suggestion by Karl Pearson, as an alternative to mid-ranks, that the ties should all be ranked as if they were the highest member of the tie.

This is subject to the obvious disadvantages that it gives different results if one ranks from the other end of the scale and that it destroys the useful property that the mean rank of the whole series shall be $\frac{1}{2}(n+1)$. So far as I know it has never been used in problems involving the calculation of ranking coefficients.

(*b*) According to Woodbury (1940), DuBois (1939), in a paper which I have been unable to consult, has suggested allotting the ties an equal rank but proposes to determine it so that the sum of squares of the ranks shall be that of an untied ranking, namely, of the first n integers, $\frac{1}{6}n(n+1)(2n+1)$. This is rather troublesome, and it is not clear to me what advantages it possesses over the mid-rank method.

3. The effect of ties on the calculation of Spearman's ρ was considered by 'Student' (1921). ρ may be regarded as the product-moment correlation between two variates given by the two rankings. Since the variance of a ranking of n is given by $\frac{1}{12}(n^2-1)$, ρ is given by

$$\rho = \frac{12}{n^3-n}\sum_{i=1}^{n}\{X_i-\tfrac{1}{2}(n+1)\}\{Y_i-\tfrac{1}{2}(n+1)\}, \tag{1}$$

where X_i and Y_i represent the two rankings. This is easily reduced to the simpler and more familiar form

$$\rho = 1-\frac{6\Sigma(d^2)}{n^3-n}, \tag{2}$$

where

$$d_i = X_i-Y_i. \tag{3}$$

Pursuing this analogy with the product-moment correlation 'Student' shows that, on the mid-rank method, the effect of a tie of t consecutive members is to lower the variance of the ranking by $\frac{1}{12n}(t^3-t)$. This is additive for any number of sets of ties in either the X- or the Y-variate, and if we write

$$T_X = \tfrac{1}{12}\Sigma(t^3-t), \tag{4}$$

the summation being over the ties of the X-variate, and T_Y for the similar sum for Y, we find for the product-moment correlation, say ρ_S,

$$\rho_S = \frac{\frac{1}{2}\operatorname{var} X + \operatorname{var} Y - \operatorname{var}(X-Y)}{\sqrt{(\operatorname{var} X \operatorname{var} Y)}}$$

$$= \frac{\frac{1}{6}(n^3-n) - (T_X+T_Y) - \Sigma(d^2)}{\sqrt{\{\frac{1}{6}(n^3-n)-2T_X\}\{\frac{1}{6}(n^3-n)-2T_Y\}}} \tag{5}$$

$$= \frac{\frac{1}{6}(n^3-n) - (T_X+T_Y) - \Sigma(d^2)}{\{\frac{1}{6}(n^3-n)-(T_X+T_Y)\}\sqrt{\left[1-\frac{(T_X-T_Y)^2}{\{\frac{1}{6}(n^3-n)-(T_X+T_Y)\}^2}\right]}}. \tag{6}$$

'Student' notes that if $T_X - T_Y$ is small, we have approximately

$$\rho_S = 1 - \frac{\Sigma(d^2)}{\frac{1}{6}(n^3-n)-(T_X+T_Y)}. \tag{7}$$

It is also useful to note that if T_X and T_Y are small compared with $\frac{1}{6}(n^3-n)$, we have

$$\rho_S = 1 - \frac{6\Sigma(d^2)}{n^3-n},$$

so that the correction to be applied to the ordinary formula is negligible for many practical purposes.

Example 1. Consider, for instance, the two rankings of 10:

X:	1	$2\frac{1}{2}$	$2\frac{1}{2}$	$4\frac{1}{2}$	$4\frac{1}{2}$	$6\frac{1}{2}$	$6\frac{1}{2}$	8	$9\frac{1}{2}$	$9\frac{1}{2}$
Y:	1	2	$4\frac{1}{2}$	$4\frac{1}{2}$	$4\frac{1}{2}$	$4\frac{1}{2}$	8	8	8	10

In the first ranking there are four tied pairs and hence

$$T_X = \tfrac{1}{12}(2^3-2) = 2.$$

In the second there is one set of four ties and one of three, and hence

$$T_Y = \tfrac{1}{12}(60+24) = 7.$$

We also have

$$\Sigma(d^2) = 13.$$

Hence, in accordance with (5),

$$\rho_S = \frac{165-22}{\sqrt{(161.151)}} = \frac{143}{155{\cdot}92}$$

$$= 0{\cdot}9171.$$

Calculation on the basis of (2) gives $\rho = 1 - \frac{78}{990} = 0{\cdot}9212.$

The value given by (7) is

$$\rho_S = 1 - \frac{13}{165-9}$$

$$= 0{\cdot}9167.$$

4. There is another way of looking at this problem which 'Student' did not mention. Suppose we regard any set t of tied ranks as due to inability on the part of the observer to distinguish real differences; i.e. we assume that there does exist a set of integral ranks although we are ignorant of it on present evidence. Then we may ask, what is the *average* value of ρ over all the $t!$ possible ways of assigning integral ranks to the tied members?

5. If the t corresponding members in the Y-ranking are held fixed, then the average covariance for all $t!$ arrangements of the X-members is the covariance of the fixed Y-members and the average of the t X-members. But this latter gives the mean ranks of the tied members, and consequently the mean covariance of the two rankings is

$$\frac{1}{n}\{\tfrac{1}{12}(n^3-n)-\tfrac{1}{2}\Sigma(d^2)-\tfrac{1}{2}(T_X+T_Y)\}, \tag{8}$$

the effect of various sets of ties being additive. If now we divide by the *actual* variances of X and Y we arrive at equation (5). Thus 'Student's' formula may be regarded as giving a mean value of the coefficient which would be obtained if the ties were replaced in all possible ways by the integral ranks which they cover; always bearing in mind that we have not averaged the variances.

6. A similar point of view has been adopted by Woodbury (1940) who does not seem to have been aware of 'Student's' results; but Woodbury takes as his variance the quantity $\frac{1}{12}(n^2-1)$, that is to say, he determines the average ρ which would be obtained if the ties were replaced by appropriate integral rankings in all possible ways, the variance in each case being that of the first n integers. This results in ρ_W, say, where

$$\rho_W = 1-\frac{6\{\Sigma(d^2)+T_X+T_Y\}}{n^3-n}, \tag{9}$$

the difference from ρ_S of equation (5) lying, of course, in the denominators in the second term on the right.

Example 2. For example, in the illustration considered above Woodbury's value would be

$$\rho_W = 1-\frac{6(13+2+7)}{990}$$
$$= 0{\cdot}8667,$$

against $\rho_s = 0{\cdot}9171$.

7. The question then arises, which is the better measure of rank correlation, ρ_S or ρ_W? It is useful in the first instance to consider some special cases.

(*a*) Suppose that the two rankings are both completely tied, i.e. that each rank is $\frac{1}{2}(n+1)$. We clearly have

$$\rho = 1,$$

indicating complete correlation. For the 'Student' form we have

$$\rho_S = \frac{0}{\sqrt{(0\times 0)}},$$

an indeterminate form which, however, may be regarded as unity as a limiting case in virtue of the next subsection. For Woodbury's form we find

$$\rho_W = \frac{0}{\frac{1}{6}(n^3-n)}$$
$$= 0,$$

indicating zero correlation. In short, Woodbury's 'correction' has reduced ρ from 1 to 0.

(*b*) Suppose that both rankings are the same, that the last member in each is ranked n and that the others are all tied and hence have rank $\frac{1}{2}n$. Then it will be found that

$$\rho = 1,$$
$$\rho_S = 1,$$
$$\rho_W = \frac{3}{n+1} \quad (n \geqslant 2).$$

The crude form of ρ and 'Student's' corrected form are in agreement that the correlation is unity. Woodbury's form differs entirely and gives a correlation which is small for large n.

(*c*) Generally, if the two rankings are identical and there are ties giving a T-number of T in each

$$\rho = \rho_S = 1,$$
$$\rho_W = 1 - \frac{12T}{n^3 - n}.$$

(*d*) Suppose that one ranking is in the natural order $1, \ldots, n$ and has no ties, so that $T_X = 0$. If the other ranking has the last member ranked n and the others completely tied we find

$$\Sigma(d^2) = \tfrac{1}{12}n(n-1)(n-2),$$
$$\rho = \frac{n+4}{2(n+1)} \sim \frac{1}{2},$$
$$\rho_S = \sqrt{\frac{3}{n+1}},$$
$$\rho_W = \frac{3}{n+1}.$$

For large n, ρ tends to 0·5, whereas ρ_S and ρ_W tend to zero, the latter faster than the former.

8. It appears to me that the decision as to which of the coefficients ρ_S or ρ_W is preferable can only rest on the use to which they are to be put.

(*a*) Let us suppose in the first instance that the objects have a definite ranking $1, \ldots, n$ determined in some objective way. The purpose of correlating the order assigned by an observer is then to determine the observer's accuracy, not the real ranking. 'Student's' form of the coefficient would measure the product-moment correlation of ranks, giving weight to the fact that if the observer produces ties the variance of his estimates is reduced. Woodbury's form would measure the average correlation of all the results obtained if the observer allotted to the tied groups integral ranks determined at random. For instance in a ranking of 8

1	2	3	4	5	6	7	8
4	4	4	4	4	4	4	8

$\rho_S = \sqrt{\frac{1}{3}} = 0{\cdot}577, \rho_W = \frac{1}{3} = 0{\cdot}333$. There does not seem to me to be much to choose between the two, but on the whole Woodbury's form gives figures nearer to what one would expect. We may suppose the observer to be in a genuine state of indecision when considering the tied members, and the average of all the values given by guessing integral ranks at random seems a fair measure of his ability. 'Student's' form gives higher values because he divides the product-moment by the actual standard deviations, and hence gives the observer credit, so to speak, for clustering his values in spite of the fact that he ought not to do so *because there really is an objective order*. In a case of this kind I should therefore favour Woodbury's form.

(*b*) The situation is quite different if no objective order is given and we are measuring the concordance between two judges. In this case Woodbury's form seems to me to give the wrong answer. In the case where two rankings are identical, for instance, one is entitled to expect that a measure of correlation should be unity—agreement could not be better. Both judges may be wrong, but that is not the point. We are measuring their agreement, not their accuracy. It has been shown above that if all members are tied Woodbury's form would give a zero correlation between the two rankings, which on the face of it seems ridiculous. We are no longer entitled to assume an objective order, or, even if there are real differences in the objects, to suppose that they fall above the threshold of the discriminatory power of the judges. 'Student's' form appears to be far better.

9. It is, of course, undeniable that 'Student's' form is more troublesome to calculate. This is unimportant if only two or three rankings are to be compared, but might be more important if there were large numbers of rankings. In such a case, however, it is more usual to work out a single measure of joint correlation rather than many pairs. The problem of m rankings for tied variates is dealt with below.

10. I proceed to consider the appropriate method of dealing with ties in calculating the alternative coefficient of correlation known as τ (Kendall, 1938). In an elegant synthesis of the rank correlation problem Daniels (1944) has recently pointed out that τ may be defined as

$$\tau = \frac{\Sigma(a_{ij}b_{ij})}{\sqrt{\{\Sigma a_{ij}^2 \Sigma b_{ij}^2\}}}, \tag{10}$$

where $$a_{ij} = -a_{ji}, \quad b_{ij} = -b_{ji},$$

a_{ij} is a score allotted to each pair of ranks X_i, X_j as $+1$ if $j > i$ and -1 if $i < j$, b_{ij} similarly relating to the Y-ranking. In the ordinary ranking case, of course,

$$\Sigma a_{ij}^2 = \Sigma b_{ij}^2 = \tfrac{1}{2}n(n-1). \tag{11}$$

Daniels' form has the advantage of revealing τ as analogous to an ordinary product-moment coefficient.

11. To extend this definition to the case of tied ranks we have only to define the score a_{ij} for equal ranks, and this is easily done by defining it as zero, midway between the scores of $+1$ and -1 taken when the ranks are unequal. This, it will be noticed, affects the denominator in the definition of equation (10) as well as the numerator.

Example 3. Let us consider the rankings of Example 1, namely,

X:	1	$2\frac{1}{2}$	$2\frac{1}{2}$	$4\frac{1}{2}$	$4\frac{1}{2}$	$6\frac{1}{2}$	$6\frac{1}{2}$	8	$9\frac{1}{2}$	$9\frac{1}{2}$
Y:	1	2	$4\frac{1}{2}$	$4\frac{1}{2}$	$4\frac{1}{2}$	$4\frac{1}{2}$	8	8	8	10

Considering the first member in association with the other 9, we see that the score in both rankings in each case is $+1$, so that the total score is 9; the second and third members in the X-ranking are tied and this pair therefore scores 0, whatever the Y-position. The score from pairs associated with the second member will be found to be 7; in the Y-ranking members 3–6 are tied and therefore the only non-vanishing scores arise from association of the third member with the seventh, eighth, ninth and tenth members, score 4. The total score will be found to be

$$9+7+4+4+4+4+3+1+1+0 = 33.$$

The sum Σa_{ij}^2 is found to be 41 and Σb_{ij}^2 is 36 and hence, writing τ_s for the corresponding quantity to ρ_s, we have

$$\tau_s = \frac{33}{\sqrt{(41 \cdot 36)}} = 0{\cdot}8589.$$

The value of ρ_s is 0·9171 but the difference need cause no concern as the coefficients do give rather different results, having different scales.

The general rule for the formation of Σa_{ij}^2 will be clear. If there is a tie of extent t we calculate $\frac{1}{12}t(t-1)$ and sum for all ties. If this sum is U then

$$\Sigma a_{ij}^2 = \tfrac{1}{12}n(n-1) - U. \tag{12}$$

12. If we replace any tied set by the corresponding integral ranks in any order and average for all the $t!$ possible orders we get the same result as by replacing a_{ij} for the tied members by zero; for in the $t!$ arrangement each pair will occur an equal number of times in the order XY and the order YX, so that the allocation of $+1$ in the first case and -1 in the second is equivalent to the allocation of zero on the average. Thus we may regard our score for tied ranks as the mean of the values obtained by allotting integral ranks in all possible ways. On the analogy of Woodbury's treatment of ρ we could then define

$$\tau_W = \frac{\text{score}}{\frac{1}{2}n(n-1)}. \tag{13}$$

The choice between τ_S and τ_W is precisely the same as in the case of Spearman's ρ; that is to say we might use the latter where an objective order is known but the former where it is a question of measuring the concordance between judges.

13. For the purposes of comparison with the special cases considered in § 7 it may be worth while giving the corresponding values of T:

(*a*) Both rankings completely tied:

$$\tau_S \text{ indeterminate}, \quad \tau_W = 0.$$

(*b*) Rankings identical and all tied except last member:

$$\tau_S = 1, \quad \tau_W = \frac{2}{n}.$$

(*c*) Rankings identical, ties giving U-member U in each:

$$\tau_S = 1, \quad \tau_W = 1 - \frac{2U}{n(n-1)}.$$

(*d*) One ranking equal to the natural order $1, \ldots, n$, the other all tied except the last member:

$$\tau_S = \sqrt{\frac{2}{n}}, \quad \tau_W = \frac{2}{n}.$$

The estimation of a ranking

14. In a previous note (1942) I considered the problem of estimating the true ranking (or the ranking on which there was the greatest measure of agreement) for m-rankings of n individual exemplified by

Object	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8
	4	2	1	7	6	3	5	8
	7	2	1	6	4	5	3	8
	7	4	2	6	5	3	1	8
Sum of ranks	18	8	4	19	15	11	9	24

It was shown that a reasonable estimate was to be obtained by ranking according to the sums of ranks, beginning with the lowest, e.g. in this example the ranking would be

$$A_3 \quad A_2 \quad A_7 \quad A_6 \quad A_5 \quad A_1 \quad A_4 \quad A_8.$$

This is the best in that it minimizes the sum of squares of deviations from what they would be if the m-rankings were identical; and it also maximizes the average ρ between the observed and the estimated rankings.

15. The above method may also be regarded in this way: if an object is ranked r, it is preferred to $n-r$ members but $r-1$ members are preferred to it. Allotting as usual $+1$ for the first type and -1 to the second we see that the individual scores $n+1-2r$ in its own ranking. Summing over the m-rankings we see that an individual ranked X_r, Y_r, Z_r, etc. has altogether a score of

$$m(n+1)-2\Sigma(X). \tag{14}$$

If then we rank the individuals according to their total scores, beginning with the highest, we arrive at exactly the same result as by ranking according to $\Sigma(X)$ beginning with the lowest. Thus our method arranges the objects in the order of numbers of preferences; a further argument in its favour. It is also easy to see that the method minimizes the sums of squares of deviations of preferences from what they would be if there were complete agreement. In fact, denote the estimated ranking by $X_1, \ldots, X_n$ and let the corresponding sums of preferences be $\xi_1, \xi_2, \ldots, \xi_n$, this being a permutation of $m(n+1)-mX_j$, $j = 1, \ldots, n$. If the actual preferences are given in sum by $S_1, \ldots, S_n$ we have to minimize

$$\sum_{j=1}^{n}(S_j-\xi_j)^2 = \Sigma S^2+\Sigma\xi^2-2\Sigma(S\xi). \tag{15}$$

The first two terms on the right are constants and we have therefore to maximize $\Sigma(S\xi)$. This is clearly done by multiplying the largest S by the largest ξ, that is to say the largest S by the smallest X, and so on. In other words, we allot X_1 to the largest S and so on in order.

16. To complete the story one would like to be able to prove that the method maximized the average τ between observed and estimated rankings.

Unfortunately the proposition fails, as is shown by the following example:

A_1	A_2	A_3	A_4
2	3	1	4
1	2	4	3
1	2	4	3
4	7	9	10

The estimated order here would be that running from left to right across the page as written, and the total score between that order and the three observed orders will be found to be $2+4+4 = 10$. But if we interchange the last two columns it becomes $0+6+6 = 12$. Such a situation, however, is of rare occurrence and can only occur when there is substantial disagreement between judges on the two objects interchanged, in which case no ranking is very reliable. I do not think, therefore, that the failure of the result in extreme cases is important.

17. Suppose now that some of the ranks are tied. Does the method of summing ranks apply unchanged to give a good estimate?

(*a*) In the first place, the method continues to give an answer which appears reasonable on the face of it. Moreover it may be regarded as an average result for all the ways of permuting the tied ranks when replaced by the appropriate integral ranks.

(*b*) If the question is regarded as one of ranking according to preferences, the replacement of a pair of integral ranks by a tie does not affect the preferences with other members and

merely cancels a preference between the tied pair; and so for any set of ties. In consequence the method preserves the property of ranking according to the number of preferences.

(*c*) If we measure the average ρ with the estimated ranking in Woodbury's form of corrected ρ the method provides a maximum average ρ unless the estimated ranking itself contains ties, in which case the result might conceivably fail, though it is unlikely to do so.

In fact, let the estimated ranking be $X_1, \ldots, X_n$ and the rank of the jth object in the kth ranking be Y_{jk}. We shall maximize the average ρ by maximizing

$$V = \sum_{k=1}^{m} \sum_{j=1}^{n} \{X_j - \tfrac{1}{2}(n+1)\}\{Y_{jk} - \tfrac{1}{2}(n+1)\}$$

$$= \sum_{j=1}^{n} \{X_i - \tfrac{1}{2}(n+1)\}\{S_j - \tfrac{1}{2}m(n+1)\}, \tag{16}$$

which reduces to maximizing $\Sigma(XS)$. If, however, there are ties in the estimated ranking our problem is to minimize something of the form

$$\frac{V}{\left\{\tfrac{1}{2}(n^2-1) - \dfrac{1}{n}\Sigma(T_X)\right\}^{\frac{1}{2}}}, \tag{17}$$

and variations in T_X may upset the result. This, however, is not likely to be serious unless there are many ties in the estimated ranking, in which case estimation of any kind is unreliable.

(*d*) If we measure the average ρ with the estimated ranking in 'Student's' form the result again may fail for (16) then becomes of the form

$$\Sigma\Sigma\{X_j - \tfrac{1}{2}(n+1)\}\, A_{jk}\{Y_{jk} - \tfrac{1}{2}(n+1)\}, \tag{18}$$

where the coefficients A_{jk} differ from our ranking to another because they depend on differing variances.

(*e*) Similar considerations apply to the proposition that the method minimizes the sums of squares of deviations from what the sums of ranks or preferences would be if all rankings were alike. Apart from complications due to ties in the estimated ranking, the minimal properties continue to obtain.

18. To sum up, therefore, the method of estimating the ranking according to sums of ranks appears to give satisfactory results when ties are involved.

Example 4. Consider the three rankings

	1	2	3	4½	4½	6	7½	7½	9	10
	1	2½	2½	4½	4½	6½	6½	8	9½	9½
	1	2	4½	4½	4½	4½	8	8	8	10
Sums of ranks	3	6½	10	13½	13½	17	22	23½	26½	29½
Estimated ranking	1	2	3	4½	4½	6	7	8	9	10

The sums of ranks give an estimated ranking as shown. There is one case here where the sums of ranks are equal and the individual ranks yielding those sums are also equal. There seems no better course than to tie them. Had the sums for these two been

4½	3½
4½	4½
4½	5½
13½	13½

we might perhaps have ranked the former as 4 and the latter as 5, on the ground that the variance of the former is less and the group therefore 'cluster' better than the other. This is a new principle deriving no support from the various minimal principles already introduced. It is the usual practice, I think, to regard an estimate as better when it is based on more closely grouped observations; but here the resulting estimate of the mean ranking is $4\frac{1}{2}$ so it can also be argued that the tie should remain. On the whole this seems to me the better course.*

TESTS OF SIGNIFICANCE FOR m RANKINGS

19. I proceed to consider what modifications, if any, are required in significance tests when ties can appear in the rankings. Babington Smith & I (1939) have discussed the case when the ranks are integral. The algebra required for the more extended discussion has been given by Pitman (1938) in considering a similar problem in the analysis of variance.

Consider an array of m rows

$$\begin{array}{cccc} a_1 & a_2 & \dots & a_n \\ b_1 & b_2 & \dots & b_n \\ \dots & \dots & \dots & \dots \\ k_1 & k_2 & \dots & k_n \end{array} \tag{19}$$

If S is the sum of square of column totals about their mean and S' the sum of squares of all values about their mean, we define

$$W = \frac{S}{mS'} \tag{20}$$

as the coefficient of concordance. In the case when the a's, b's, etc. are permutations of the natural numbers 1 to n we have

$$W = \frac{S}{\frac{1}{12}m^2(n^3-n)}. \tag{21}$$

W can vary from 0 to 1, attaining the latter value only if all rankings are identical.

20. Let us in the first place consider what happens to formula (21) when some of the integral ranks are replaced by ties. If the T-numbers for the various rows are T_a, T_b, etc., the formula becomes

$$W = \frac{S}{\frac{1}{12}m^2(n^3-n)-m\Sigma(T)}. \tag{22}$$

This is as simple a form as we require.

In the data of Example 4, for instance, we find

$$S = 682{\cdot}9,$$

$$W = \frac{682{\cdot}9}{742{\cdot}5-30}$$

$$= 0{\cdot}958.$$

* In my note (1942), which dealt only with integral ranks, I suggested that, where the sums of ranks are equal, precedence should be given to the one with the smaller variance; but I was there considering only an estimated ranking which itself was integral. When ties are permitted I should, as stated above, use them where sums of ranks are equal.

21. Denoting by α_r the rth moment of the a-row in (19), and similarly for β_r, γ_r, etc., and by α'_r the rth k-statistic, we have for the moments of W, from Pitman's results,

$$\overline{W} = E(W) = \frac{1}{m}, \tag{23}$$

$$E(W-\overline{W})^2 = \frac{4}{m^2(n-1)}\frac{\Sigma\alpha_2\beta^2}{\Sigma^2\alpha_2}, \tag{24}$$

$$E(W-\overline{W})^3 = \frac{48}{m^3(n-1)}\frac{\Sigma\alpha_2\beta_2\gamma_2}{\Sigma^3\alpha_2} + \frac{8(n-1)(n-2)}{m^3n^4}\frac{\Sigma\alpha'_2\beta'_2}{\Sigma^3\alpha_2}, \tag{25}$$

$$E(W-\overline{W})^4 = \frac{48}{m^4(n-1)^2}\frac{\Sigma^2\alpha_2\beta_2}{\Sigma^4\alpha_2} - \frac{96}{m^4(n-1)^2(n+1)}\frac{\Sigma\alpha_2^2\beta_2^2}{\Sigma^4\alpha_2} + \frac{1152}{m^4(n-1)^3}\frac{\Sigma\alpha_2\beta_2\gamma_2\delta_2}{\Sigma^4\alpha_2}$$
$$+ \frac{16(n-1)(n-2)(n-3)}{m^4n^5(n+1)}\frac{\Sigma\alpha'_4\beta'_4}{\Sigma^4\alpha_2} + \frac{252(n-2)}{m^4n^4}\frac{\Sigma\alpha'_3\beta'_3\gamma_2}{\Sigma^4\alpha_2}. \tag{26}$$

In the case when the numbers are permutations of the first n integers these expressions reduce to

$$\overline{W} = \frac{1}{m}, \tag{27}$$

$$E(W-\overline{W})^2 = \frac{2(m-1)}{m^3(n-1)}, \tag{28}$$

$$E(W-\overline{W})^3 = \frac{8(m-1)(m-2)}{m^5(n-1)^2}, \tag{29}$$

$$E(W-\overline{W})^4 = \frac{12(m-1)^2}{m^6(n-1)^2} + \frac{48(m-1)(m-2)(m-3)}{m^7(n-1)^3} - \frac{48(m-1)}{m^7(n-1)^2(n+1)}. \tag{30}$$

If m and n are moderately large, these expressions are approximately the same (exactly so for the first two) as the moments of

$$dF = \frac{1}{B(p,q)}W^{p-1}(1-W)^{q-1}dW, \tag{31}$$

where

$$\left.\begin{aligned} p &= \tfrac{1}{2}(n-1)-\frac{1}{m}, \\ q &= (m-1)p. \end{aligned}\right\} \tag{32}$$

It follows that W can be tested in Fisher's z-distribution by writing

$$z = \tfrac{1}{2}\log\frac{(m-1)W}{1-W}, \tag{33}$$

$$\left.\begin{aligned} \nu_1 &= (n-1)-\frac{2}{m}, \\ \nu_2 &= (m-1)\nu_1. \end{aligned}\right\} \tag{34}$$

How far does this require modification for tied ranks?

22. For the purposes of an accurate test we can, of course, work out the first four moments of W from (23) to (26) in individual cases and fit an *ad hoc* curve; but this is a tedious process and some approximation is necessary.

The first two moments of (31) are

$$\mu_1 \text{ (about zero)} = \frac{p}{p+q},$$

$$\mu_2 = \frac{pq}{(p+q)^2(p+q+1)},$$

and if we identify them with the first two moments of W, $\frac{1}{m}$ and $\mu_2(W)$, say, we find

$$\left.\begin{aligned} p &= -\frac{1}{m} + \frac{m-1}{m^3\mu_2(W)}, \\ q &= (m-1)\,p, \end{aligned}\right\} \tag{35}$$

so that approximately W can be tested in the z-distribution with

$$\left.\begin{aligned} \nu_1 &= -\frac{2}{m} + \frac{2(m-1)}{m^3\mu_2(W)}, \\ \nu_2 &= (m-1)\,\nu_1. \end{aligned}\right\} \tag{36}$$

23. We have, as in (24),

$$\mu_2(W) = \frac{4}{m^2(n-1)} \frac{\Sigma\alpha_2\beta_2}{\Sigma^2\alpha_2}.$$

Writing A for $\Sigma\alpha_2$ and B for $\Sigma\alpha_2^2$ we have

$$\mu_2(W) = \frac{2}{m^2(n-1)} \left\{1 - \frac{B}{A^2}\right\},$$

so that the appropriate degrees of freedom are

$$\left.\begin{aligned} \nu_1 &= -\frac{2}{m} + \frac{(n-1)(m-1)}{m(1-B/A^2)}, \\ \nu_2 &= (m-1)\,\nu_1. \end{aligned}\right\} \tag{37}$$

If the T-numbers appropriate to the various rankings are small compared with $\frac{1}{12}(n^3-n)$ we can approximate further. In fact write N for $\frac{1}{12}(n^3-n)$. Then

$$A = mN - \Sigma(T),$$

$$B = mN^2 - 2N\Sigma(T) + \Sigma(T^2),$$

and to the first order in $\Sigma(T)$

$$\begin{aligned} \frac{B}{A^2} &= \frac{mN^2 - 2N\Sigma(T)}{m^2N^2 - 2mN\Sigma(T)} \\ &= \frac{1}{m}\left\{1 - \frac{2\Sigma(T)}{mN}\right\}\left\{1 - \frac{2\Sigma(T)}{mN}\right\}^{-1} \\ &= \frac{1}{m}. \end{aligned}$$

On substitution in (27)

$$\left.\begin{aligned} \nu_1 &= -\frac{2}{m} + (n-1), \\ \nu_2 &= (m-1)\,\nu_1. \end{aligned}\right\} \tag{38}$$

This, of course, is the same as (34) so that, if the number or extent of the ties is small, the test for untied ranks requires no modification (other than that necessary in the calculation of W itself).

24. It thus appears that we can apply the usual test unchanged unless the ties are numerous enough to render $\Sigma(T)$ not small compared with mN. If the ties are numerous we can work with the modified degrees of freedom given by (35), but in such a case it would probably be as well to verify by direct calculation that the third and fourth moments of W were in reasonable agreement with those given by the β-approximation implicit in the use of the z-test. If it happens that one or two rankings contribute the major part of $\Sigma(T)$ we may perhaps reject them on the grounds that the judges are very bad, but the rejection of observations has to be done with some care and only after we are satisfied that they really are exceptional and not merely outlying members of a continuous range.

Addition of extra members to a ranking

25. There is one class of case in which I have found the coefficient τ to have definite advantages over ρ. An example will illustrate the point best. Suppose I send out an inquiry to a number of firms asking for some information which they may or may not wish to disclose; and suppose that the information is of a type for which one would expect that the non-participants might differ from the participants. By a certain date a number of replies have been received and it is then necessary to close the inquiry and to summarize the results. How far can I assume that the replies to hand are representative of the population to which the inquiry was addressed? Is there any evidence to suggest that those who reply earlier to the inquiry differ from those who reply later?

26. To simplify the illustration suppose that I receive 15 replies in the form of a percentage figure which occur in the following order:

Order of receipt:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Percentage:	15	13	12	16	25	8	9	14	17	11	18	20	10	21	19
Rank of percentage:	8	6	5	9	15	1	2	7	10	4	11	13	3	14	12

I have chosen percentages which are all different so as not to complicate the example, but if ties are present there is no essential modification.

Now if there is some relation between the order of reply and the magnitude of the percentage, it ought to be shown up by the rank correlation between the order of reply and the order of magnitude of the percentage. The latter is shown in the last row of the example above and we find

$$\Sigma(d^2) = 392,$$

$$\rho = 0{\cdot}300.$$

This in fact, is barely significant, but I am not for the moment concerned with significance. Suppose that after we have completed this calculation two more replies arrive with percentages 7 and 23. We now have to calculate a revised value of ρ by re-numbering nearly all the replies and working *ab initio*. In practice the continual arrival of stragglers is quite common and to work out a new value of ρ each time is a great arithmetical nuisance. The point I wish to make is that τ is not subject to this disability, extra values being capable of addition as required.

In the above example for 15 members the value of $\Sigma(a_{ij}b_{ij})$ is easily seen to be

$$0+3+4+1-10+9+8+3+2+3+2+1+2-1 = 27,$$

so that

$$\tau = \tfrac{27}{105} = 0{\cdot}257.$$

If now we add a 16th member with the value 7, the contribution to $\Sigma(ab)$ is obtained by considering this new member in conjunction with the other fifteen, and is seen to be -15. Similarly, a further member valued 23 adds 13. The new value of $\Sigma(ab)$ is thus 25 and the new τ is given by

$$\tau = \tfrac{25}{136} = 0{\cdot}184.$$

In this way a kind of running value of τ can be ascertained without re-ranking at each stage as is necessary with ρ. Thus τ has a decided advantage in this class of case, namely the calculation of ranking coefficients for time series which may be extended in length.

REFERENCES

DANIELS, H. E. (1944). The relation between measures of correlation in the universe of sample permutations. *Biometrika*, **33**, 129.

DUBOIS, P. (1939). Formulas and tables for rank correlation. *Psychol. Rec.* **3**, 45.

KENDALL, M. G. (1938). A new measure of rank correlation. *Biometrika*, **30**, 81.

KENDALL, M. G. (1942). Note on the estimation of a ranking. *J.R. Statist. Soc.* **105**, 119.

KENDALL, M. G. & BABINGTON SMITH, B. (1939). The problem of m rankings. *Ann. Math. Statist.* **10**, 275.

PITMAN, E. J. G. (1938). Significance tests which may be applied to samples from any population. III. The analysis of variance test. *Biometrika*, **29**, 322.

'STUDENT' (1921). An experimental determination of the probable error of Dr Spearman's correlation coefficient. *Biometrika*, **13**, 263.

WOODBURY, M. A. (1940). Rank correlation when there are equal variates. *Ann. Math. Statist.* **11**, 358.

THE PROBABILITY INTEGRAL OF THE MEAN DEVIATION

EDITORIAL NOTE

1. About 3 years ago a need arose to obtain the probability levels of the mean deviation in random samples from a normal population. The requirement was in a field of production quality control where it was customary, as well as convenient, to use the mean deviation as a measure of dispersion in a sample, rather than the standard deviation or the range between extreme individuals. The need was pressing, and it appeared that the quickest answer for practical purposes would be obtained by using the known expressions for the mean and variance of the M.D. and getting a measure of the departure of the distributions from normality by a sampling experiment with random numbers. An investigation on these lines was undertaken by Dr E. H. Sealy and Mr C. D. Bates of the Advisory Service on Quality Control, Ministry of Supply. Their results, in the form of a table of factors for control limits for sample sizes varying from $n = 5$ to 20, were issued in 1943.

2. These limits, though adequate for the immediate requirement, were of course not exact and are now superseded in the range $n = 2$ to 10 by the tables printed below in Mr Godwin's paper.

3. At the time when the planning of the earlier investigation was discussed with Dr Sealy, I had overlooked the fact that R. C. Geary's paper of 1936 on the distribution of the ratio of the M.D. to the S.D. in samples from a normal population contained formulae from which the higher moments of the M.D. could be derived. When this oversight was realized and before Mr Godwin's work was undertaken, I had asked Dr Geary to develop from his earlier work, expansions for the 3rd and 4th order semi-invariants of the M.D. in terms of inverse powers of $\nu = n - 1$ (where n is the sample size). It seems of interest to take this opportunity of putting these results on record.

4. *Dr Geary's expansions* (population standard deviation as unit).

Mean deviation in a sample of n observations:

$$m = \frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}|. \tag{1}$$

Expectation of m:

$$\overline{m} = \sqrt{\left\{\frac{2(n-1)}{\pi n}\right\}}. \tag{2}$$

Semi-invariants of the sampling distribution of m when the population is normal:

$$\lambda_3 = \left(\frac{\nu}{\nu+1}\right)^{\frac{1}{2}}\left\{\frac{0{\cdot}218\,014}{\nu^2} - \frac{0{\cdot}074\,170}{\nu^3} + \frac{0{\cdot}057\,313}{\nu^4} - \frac{0{\cdot}040\,457}{\nu^5} + \frac{0{\cdot}023\,601}{\nu^6} - \ldots\right\}, \tag{3}$$

$$\lambda_4 = \left(\frac{\nu}{\nu+1}\right)^{2}\left\{\frac{0{\cdot}114\,771}{\nu^3} - \frac{0{\cdot}068\,509}{\nu^4} + \frac{0{\cdot}033\,371}{\nu^5} - \frac{0{\cdot}120\,003}{\nu^6} + \ldots\right\}. \tag{4}$$

The frequency constants β_1 and β_2 can be obtained from

$$\beta_1 = \lambda_3^2/\lambda_2^3, \quad \beta_2 = 3 + \lambda_4/\lambda_2^2, \tag{5}$$

where

$$\lambda_2 = \sigma_m^2 = \frac{2(n-1)}{n^2\pi}\left\{\tfrac{1}{2}\pi + \sqrt{\{n(n-2)\}} - n + \sin^{-1}\frac{1}{n-1}\right\}. \tag{6}$$

It should be noted that Geary (1936) has taken n' for the sample size and written $n = n' - 1$, but to be consistent with the paper which follows, I have written n for his n' and ν for his $n = n' - 1$.*

5. The accompanying table shows the 3rd and 4th semi-invariants of m computed from the expansions given above and also the moment ratios β_1 and β_2; for $n = 4$ the figures for β_1 and β_2 may be compared with R. A. Fisher's (1920) values of $\beta_1 = 0{\cdot}297$, $\beta_2 = 3{\cdot}280$. Differences between the expansion and true values will become rapidly less as n increases.

The sampling moments of the mean deviation

Sample size n	λ_3	λ_4	β_1	β_2
4	0·014 32	0·001 900	0·299	3·244
5	0·009 057	0·000 9785	0·230	3·194
6	0·006 244	0·000 5636	0·187	3·160
8	0·003 483	0·000 2351	0·136	3·118
10	0·002 218	0·000 1193	0·106	3·093
12	0·001 536	0·000 06864	0·088	3·076
15	0·000 9800	0·000 03492	0·069	3·060
20	0·000 5495	0·000 01461	0·051	3·045

REFERENCES

FISHER, R. A. (1920). *Mon. Not. R. Astr. Soc.* **80**, 8.
GEARY, R. C. (1936). *Biometrika*, **28**, 295.
SEALY, E. H. & BATES, C. D. (1943). Ministry of Supply, Advisory Service on Quality Control. *Technical Report*, QC/R/1.

* In equations (8) and (22) of Geary's paper (pp. 296 and 300), for d read $d' = \sum_{i=1}^{n'} |x_i - \bar{x}| / \sqrt{\{n'(n'-1)\}}$.

E.S.P.

ON THE DISTRIBUTION OF THE ESTIMATE OF MEAN DEVIATION OBTAINED FROM SAMPLES FROM A NORMAL POPULATION

By H. J. GODWIN, *Advisory Service on Statistical Method, Ministry of Supply*

The relative merits and demerits of the mean deviation and standard deviation as measures of the dispersion of a population have been discussed by Fisher (1920): though the balance is rather in favour of the latter, the mean deviation is widely used, especially in experimental work where many small samples are taken and where saving in computation is a consideration. The distribution of the estimate given by random samples from a Normal population has not previously been obtained, save for the special cases of sample sizes four (by Fisher, 1920) and five (by Jones): Helmert (1876), and later Fisher, found the second moment of the distribution, and Geary (1936) found the third and fourth moments and showed that β_1 and β_2 were $O(n^{-1})$ and $O(n^{-1})$ respectively for large sample size n. Thus the distribution may be approximated to by a Normal distribution, and this approximation improves as n increases. The distribution was estimated empirically, for small sample sizes, by Sealy & Bates (1943). In the present paper an expression for the distribution for general n is found, suitable for calculation by quadratures and Table 1 gives the resulting probability integral to 5 decimal places for $n = 2$ to 10. From this table certain percentage points have been calculated and are given in Table 2. The Normal approximations to the percentage points for sample size ten are given for comparison with the true values: Sealy's values were much closer, being least good for the extreme percentage points of the smallest sample sizes.

Let the sample values be $x_1, x_2, \ldots, x_n$, where the x's are distributed according to the frequency function $\frac{1}{\sqrt{(2\pi)}} e^{-\frac{1}{2}x^2} dx$ (there is no loss of generality in taking the population mean and standard deviation to be zero and unity respectively). The numbers $1, 2, \ldots, n$ can be assigned to the members of the sample in $n!$ ways: we suppose that $x_1 \leqslant x_2 \leqslant x_3 \ldots \leqslant x_n$. We consider separately the cases when the mean falls between x_1 and x_2, x_2 and x_3, ..., x_{n-1} and x_n. Suppose $x_k \leqslant \frac{\Sigma x_i}{n} \leqslant x_{k+1}$.

Then the mean deviation

$$m = \frac{(x_{k+1} + \ldots + x_n) - (x_1 + \ldots + x_k) - \frac{n-2k}{n} \Sigma x_i}{n}.$$

i.e.
$$\frac{n^2 m}{2} = k(x_{k+1} + \ldots + x_n) - (n-k)(x_1 + \ldots + x_k).$$

The frequency function of m is found by evaluating

$$(2\pi)^{-\frac{1}{2}n} \int \ldots \int e^{-\frac{1}{2}(\Sigma x_i^2)} dx_1 \ldots dx_n \tag{1}$$

over the region defined by
$$x_1 \leqslant x_2 \leqslant \ldots \leqslant x_n, \tag{2}$$

$$x_k \leqslant \frac{\Sigma x_i}{n} \leqslant x_{k+1}, \tag{3}$$

and
$$\frac{n^2 m}{2} \leqslant k(x_{k+1} + \ldots + x_n) - (n-k)(x_1 + \ldots + x_k) \leqslant \frac{n^2(m+dm)}{2}. \tag{4}$$

The various functions so obtained are summed over k from 1 to $n-1$, and the whole multiplied by $n!$. The integral is evaluated by a transformation of the quadratic form Σx_i^2: this is most easily done in two stages.

First put $$x_{i+1}-x_i = y_i.$$

Then (2) becomes $$y_i \geqslant 0, \tag{5}$$

(3) becomes $$y_1+2y_2+\ldots+(k-1)y_{k-1} \leqslant (n-k)y_k+\ldots+y_{n-1}$$

and $$y_1+2y_2+\ldots+ky_k \leqslant (n-k-1)y_{k+1}+\ldots+y_{n-1} \tag{6}$$

and (4) becomes

$$\frac{n^2m}{2} \leqslant (n-k)y_1+2(n-k)y_2+\ldots+k(n-k)y_k+k(n-k-1)y_{k+1}+\ldots+ky_{n-1} \leqslant \frac{n^2(m+dm)}{2}. \tag{7}$$

Now put $$u_j = y_1+2y_2+\ldots+jy_j \quad (j \leqslant k-1),$$

$$u_j = (n-j-1)y_{j+1}+\ldots+y_{n-1} \quad (j \geqslant k).$$

Then (5), (6) and (7) become

$$O \leqslant u_1 \leqslant u_2 \leqslant \ldots \leqslant u_{k-1} \leqslant \frac{n(m+dm)}{2},$$

and $$O \leqslant u_{n-2} \leqslant u_{n-3} \leqslant \ldots \leqslant u_k \leqslant \frac{n(m+dm)}{2}, \tag{8}$$

$$\frac{\partial(x_1, x_2, \ldots, x_n)}{\partial(x_1, m, n, \ldots, u_{n-2})} = \frac{n^2}{2k!(n-k)!},$$

and $$\Sigma x_i^2 = nx_1^2+\sum_{j=1}^{n-1}2x_1(n-j)y_j+\sum_{j=1}^{n-1}(n-j)y_j^2+2\sum_{j=1}^{n-2}\sum_{l=j+1}^{n-1}(n-l)y_jy_l$$

$$= n\left(x_1+\sum_{j=1}^{k-1}\frac{u_j}{j(j+1)}+\frac{nm}{2k}\right)^2+\sum_1^{k-1}\frac{u_j^2}{j(j+1)}+\sum_k^{n-2}\frac{u_j^2}{(n-j)(n-j-1)}+\frac{m^2n^3}{4k(n-k)}.$$

We now define a series of functions $G_r(x)$, such that

$$G_0(x) = 1, \quad G_r(x) = \int_0^x \exp\left[-\frac{t^2}{2r(r+1)}\right]G_{r-1}(t)\,dt. \tag{9}$$

The integral (1) now appears as the product of n simple integrals and, after integration subject to the restrictions (8), gives

$$\frac{n^2}{2k!(n-k)!}(2\pi)^{-\frac{1}{2}n}\frac{\sqrt{(2\pi)}}{\sqrt{n}}\exp\left[-\frac{m^2n^3}{8k(n-k)}\right]G_{k-1}\left(\frac{nm}{2}\right)G_{n-k-1}\left(\frac{nm}{2}\right)dm.$$

The frequency function of m is accordingly

$$f_n(m)\,dm = n!\sum_{k=1}^{n-1}\frac{n^2}{2k!(n-k)!}(2\pi)^{-\frac{1}{2}n}\frac{\sqrt{(2\pi)}}{\sqrt{n}}\exp\left[-\frac{m^2n^3}{8k(n-k)}\right]G_{k-1}\left(\frac{nm}{2}\right)G_{n-k-1}\left(\frac{nm}{2}\right)dm$$

$$= \frac{n^{\frac{1}{2}}}{2^{\frac{1}{2}(n+1)}\pi^{\frac{1}{2}(n-1)}}\left\{\sum_{k=1}^{n-1}{}^nC_k\exp\left[-\frac{m^2n^3}{8k(n-k)}\right]G_{k-1}\left(\frac{nm}{2}\right)G_{n-k-1}\left(\frac{nm}{2}\right)\right\}dm. \tag{10}$$

The calculation of the G-functions, the distribution function of m, and the percentage points was done under the direction of Dr H. O. Hartley, whose care and assistance I gratefully acknowledge: a note by him on the method of computation appears below as an Appendix.

Although the theoretical part of the work was done privately, the computations were carried out, and the work prepared for publication, under the auspices of the Ministry of Supply (S.R. 17) from whom permission to publish has been obtained.

REFERENCES

FISHER, R. A. (1920). *Mon. Not. R. Astr. Soc.* **80**, 8, June 1920.
GEARY, R. C. (1936). *Biometrika*, **28**, 295–305.
HELMERT, R. F. (1876). *Astr. Nachr.* **88**, no. 2096.
JONES, A. E. Unpublished paper.
SEALY, E. H. & BATES, C. D. (1947). Ministry of Supply, Advisory Service on Quality Control. *Technical Report*, QC/R/1.

APPENDIX

NOTE ON THE CALCULATION OF THE DISTRIBUTION OF THE ESTIMATE OF MEAN DEVIATION IN NORMAL SAMPLES

By H. O. HARTLEY, *Scientific Computing Service, Ltd.*

The formula used for the computation of this distribution function is the finite series (10) of the above paper. This expression involves the functions $G_r(x)$ which are defined by the recurrence formula (9). The numerical work therefore consists of:

(*a*) The calculation of the $G_r(x)$ by a recurrence of numerical quadratures.

(*b*) The calculation of the distribution functions $f_n(m)$ from formula (10).

(*c*) The numerical quadratures $\int_0^m f_n(m)\,dm$ yielding, for each n, the probability integral for the mean deviation m.

(*a*) The essential feature of the numerical quadratures is a new method on the National Accounting Machine. With this method, of which it is hoped to publish details in due course, it is possible to produce a table of the integral $\int_a^x f(x)\,dx$ from the 4th differences of the integrand $f(x)$ in a single operation.

The starting point of the recurrence was the table of

$$\frac{1}{\sqrt{\pi}}G_1(x) = \frac{1}{\sqrt{\pi}}\int_0^x e^{-\frac{1}{2}t^2}\,dt = \frac{2}{\sqrt{\pi}}\int_0^{\frac{1}{2}x} e^{-\alpha^2}\,d\alpha,$$

given in *Tables of the Probability Integral*, vol. **1**, W.P.A., New York.

This integral was multiplied by the ordinate

$$\frac{2}{\sqrt{\pi}}e^{-\frac{1}{2}x^2} = \frac{2}{\sqrt{\pi}}e^{-(\frac{1}{2}x)^2},$$

which is tabulated next to it in the W.P.A. Tables. Products were formed at interval 0·05 in x and these were differenced. The function*

$$(480)\,\pi^{-1}\,2^{-\frac{1}{2}}\,G_2(x)$$

was then obtained by the mechanical method of numerical quadrature in accordance with formula (9).

This process was then repeated for $r = 3, \ldots, 8$; producing the functions $G_r(x)$ with increasing constant factors and for ranges as shown below:

r	Factors of $G_r(x)$	Range
2	$480\ \pi^{-1}\,2^{-\frac{1}{2}}$	$x = 0\ (0{\cdot}05)\ 13{\cdot}5$
3	$480^2\,\pi^{-\frac{3}{2}}\,2^{-1}$	$x = 0\ (0{\cdot}05)\ 19{\cdot}0$
4	$480^3\,\pi^{-2}\,2^{-\frac{3}{2}}$	$x = 0\ (0{\cdot}05)\ 25{\cdot}0$
5	$480^4\,\pi^{-\frac{5}{2}}\,2^{-3}$	$x = 0\ (0{\cdot}1)\ 31{\cdot}0$
6	$480^5\,\pi^{-3}\,2^{-\frac{5}{2}}$	$x = 0\ (0{\cdot}1)\ 36{\cdot}0$
7	$480^6\,\pi^{-\frac{7}{2}}\,2^{-4}$	$x = 0\ (0{\cdot}1)\ 39{\cdot}0$
8	$480^7\,\pi^{-4}\,2^{-\frac{7}{2}}$	$x = 0\ (0{\cdot}1)\ 46{\cdot}0$

* The factor (480) is necessitated by the method of quadrature. The same applies to those shown in the table below.

Seven significant figures were accurate in the maximum value of G_2, but this accuracy gradually decreased to 5 significant figures for the maximum value of G_8. The ordinate functions

$$\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{t^2}{2r(r+1)}\right]$$

which occur as multipliers in formula (9) were obtained by interpolation in the Tables of z (Table II, *Tables for Statisticians and Biometricians*, vol. 1).

(*b*) For convenience of computation formula (10) was rewritten as follows:

$$f_n(m) = n!\, 2^{-\frac{1}{2}(n+1)}\pi^{-\frac{1}{2}(n-1)}\sum_{k=1}^{n} {}^nC_k\, g_{k-1}(x)\, g_{n-k-1}(x), \qquad (11)$$

where $x = \frac{1}{2}nm$ and $g_r(x) = G_r(x)\exp\left[-\frac{x^2}{2(r+1)}\right]$.

Using the symmetry in k the number of terms may be halved.

The G-functions were first converted to g-functions through multiplication by the ordinates

$$\frac{c_r}{\sqrt{2\pi}}\exp\left[-\frac{x^2}{2(r+1)}\right] \quad (c_r = \text{suitably chosen constant}).$$

These ordinates were obtained by interpolation in the z-tables of Table II of *Tables for Statisticians and Biometricians*, vol. 1. The x-interval was 0·05 for $g_0, \ldots, g_4$ and 0·2 for $g_5, \ldots, g_8$. Formula (11) was then applied to obtain $f_n(m)$ at the following x- and m-intervals:

r	3	4	5	6	7	8	9	10
x-interval	0·075	0·10	0·125	0·15	0·175	0·2	0·225	0·25
m-interval	0·05	0·05	0·05	0·05	0·05	0·05	0·05	0·05

In certain cases it will be seen that the x-interval is not a tabular interval for the g-functions. In such cases Lagrangian Interpolation had to be applied. Finally the $f_n(m)$ functions were subtabulated to the final m-interval of 0·01.

(*c*) The published tables of the probability integrals $\int_0^m f_n(m)\,dm$ were then obtained by the process of mechanical quadrature on the National Machine. The tables of percentage points were obtained by inverse interpolation.

Checks were as follows. Apart from the usual checks by differencing, the first and second moments μ_1' and μ_2' were calculated from the final tables of $f_n(m)$ as well as $\int_0^m f_n(m)\,dm$. These were compared with check values calculated from the theoretical formulae

$$\mu_1' = \left(\frac{2(n-1)}{n\pi}\right)^{\frac{1}{2}},$$

$$\mu_2' = \frac{2(n-1)}{n^2\pi}\left\{2\tan^{-1}\left(\frac{n}{n-2}\right)^{\frac{1}{2}} + n^{\frac{1}{2}}(n-2)^{\frac{1}{2}}\right\}.$$

Five-decimal agreement was obtained throughout.

Grateful acknowledgement is made to Mr M. Sumner for the expert help rendered in the calculation of these tables.

TABLES OF THE
PROBABILITY INTEGRAL OF THE
MEAN DEVIATION IN NORMAL SAMPLES

Table 1. *The probability integral of the mean deviation (m) in normal samples of n observations.* (*Population standard deviation as unit*)

m \ n	2	3	4	5	6	7	8	9	10
0·00	0·00000	0·00000							
·01	·01128	·00019	0·00000						
·02	·02256	·00074	·00003	0·00000					
·03	·03384	·00167	·00009	·00001					
·04	·04511	·00297	·00022	·00002					
0·05	0·05637	0·00464	0·00042	0·00004	0·00000				
·06	·06762	·00668	·00073	·00009	·00001				
·07	·07886	·00908	·00115	·00016	·00002	0·00000			
·08	·09008	·01184	·00172	·00027	·00004	·00001			
·09	·10128	·01496	·00244	·00042	·00007	·00001			
0·10	0·11246	0·01843	0·00333	0·00064	0·00012	0·00002	0·00000		
·11	·12362	·02226	·00442	·00093	·00019	·00004	·00001		
·12	·13476	·02644	·00571	·00130	·00030	·00007	·00002	0·00000	
·13	·14587	·03095	·00723	·00178	·00044	·00011	·00003	·00001	
·14	·15695	·03581	·00899	·00237	·00064	·00017	·00005	·00001	0·00000
0·15	0·16800	0·04100	0·01101	0·00310	0·00089	0·00026	0·00008	0·00002	0·00001
·16	·17901	·04651	·01329	·00398	·00122	·00037	·00012	·00004	·00001
·17	·18999	·05234	·01585	·00503	·00163	·00053	·00018	·00006	·00002
·18	·20094	·05849	·01871	·00626	·00214	·00074	·00026	·00009	·00003
·19	·21184	·06495	·02187	·00770	·00277	·00101	·00038	·00014	·00005
0·20	0·22270	0·07171	0·02534	0·00936	0·00354	0·00135	0·00053	0·00021	0·00008
·21	·23352	·07876	·02914	·01126	·00445	·00179	·00073	·00030	·00012
·22	·24430	·08610	·03327	·01342	·00554	·00232	·00099	·00042	·00018
·23	·25502	·09371	·03773	·01585	·00682	·00297	·00132	·00059	·00026
·24	·26570	·10160	·04254	·01858	·00830	·00377	·00173	·00080	·00037
0·25	0·27633	0·10974	0·04769	0·02161	0·01002	0·00472	0·00225	0·00108	0·00052
·26	·28690	·11814	·05320	·02497	·01199	·00585	·00289	·00143	·00072
·27	·29742	·12679	·05907	·02867	·01423	·00717	·00366	·00188	·00097
·28	·30788	·13567	·06528	·03271	·01677	·00873	·00459	·00244	·00130
·29	·31828	·14478	·07186	·03713	·01962	·01052	·00571	·00312	·00172
0·30	0·32863	0·15410	0·07879	0·04192	0·02280	0·01259	0·00703	0·00395	0·00224
·31	·33891	·16364	·08608	·04709	·02634	·01495	·00858	·00496	·00289
·32	·34913	·17337	·09372	·05266	·03025	·01763	·01039	·00616	·00368
·33	·35928	·18330	·10171	·05864	·03455	·02065	·01248	·00759	·00465
·34	·36936	·19340	·11005	·06503	·03926	·02404	·01488	·00928	·00582
0·35	0·37938	0·20367	0·11872	0·07183	0·04439	0·02783	0·01762	0·01124	0·00722
·36	·38933	·21410	·12773	·07905	·04996	·03202	·02073	·01352	·00887
·37	·39921	·22469	·13706	·08670	·05598	·03666	·02424	·01615	·01082
·38	·40901	·23541	·14671	·09476	·06247	·04175	·02817	·01915	·01309
·39	·41874	·24626	·15667	·10325	·06942	·04731	·03256	·02257	·01573
0·40	0·42839	0·25724	0·16693	0·11215	0·07686	0·05338	0·03742	0·02643	0·01877
·41	·43797	·26832	·17748	·12147	·08478	·05995	·04279	·03076	·02224
·42	·44747	·27951	·18831	·13120	·09319	·06705	·04869	·03560	·02618
·43	·45689	·29079	·19941	·14133	·10209	·07468	·05513	·04099	·03063
·44	·46623	·30215	·21075	·15186	·11148	·08286	·06215	·04693	·03563
0·45	0·47548	0·31358	0·22234	0·16277	0·12136	0·09160	0·06975	0·05347	0·04121
·46	·48466	·32507	·23416	·17405	·13172	·10089	·07796	·06063	·04741
·47	·49375	·33661	·24620	·18569	·14255	·11074	·08677	·06843	·05425
·48	·50275	·34820	·25843	·19768	·15386	·12115	·09621	·07689	·06177
·49	·51167	·35982	·27086	·21001	·16562	·13212	·10627	·08602	·06998
0·50	0·52050	0·37146	0·28345	0·22265	0·17783	0·14364	0·11697	0·09584	0·07892

Table 1 (cont.). *The probability integral of the mean deviation (m) in normal samples of n observations.* (*Population standard deviation as unit*)

m \ n	2	3	4	5	6	7	8	9	10
0·50	0·52650	0·37146	0·28345	0·22265	0·17783	0·14364	0·11697	0·09584	0·07892
·51	·52924	·38313	·29620	·23559	·19047	·15569	·12829	·10635	·08860
·52	·53790	·39479	·30909	·24881	·20352	·16828	·14023	·11756	·09903
·53	·54646	·40646	·32211	·26230	·21697	·18137	·15279	·12947	·11022
·54	·55494	·41811	·33525	·27604	·23079	·19497	·16595	·14207	·12218
0·55	0·56332	0·42975	0·34847	0·28999	0·24497	0·20904	0·17970	0·15536	0·13491
·56	·57162	·44135	·36178	·30416	·25948	·22357	·19402	·16931	·14840
·57	·57982	·45293	·37516	·31851	·27430	·23852	·20889	·18392	·16264
·58	·58792	·46446	·38858	·33302	·28941	·25389	·22427	·19916	·17761
·59	·59594	·47593	·40204	·34768	·30477	·26963	·24016	·21501	·19329
0·60	0·60386	0·48735	0·41552	0·36245	0·32037	0·28572	0·25650	0·23144	0·20965
·61	·61168	·49871	·42901	·37733	·33617	·30213	·27328	·24840	·22667
·62	·61941	·50999	·44249	·39228	·35215	·31882	·29046	·26588	·24431
·63	·62705	·52120	·45594	·40729	·36827	·33576	·30799	·28383	·26252
·64	·63459	·53232	·46936	·42233	·38452	·35293	·32585	·30220	·28127
0·65	0·64203	0·54335	0·48273	0·43739	0·40087	0·37027	0·34398	0·32095	0·30050
·66	·64938	·55428	·49603	·45244	·41727	·38777	·36235	·34004	·32016
·67	·65663	·56511	·50926	·46746	·43372	·40537	·38092	·35941	·34021
·68	·66378	·57583	·52240	·48244	·45017	·42306	·39965	·37902	·36058
·69	·67084	·58644	·53544	·49734	·46661	·44078	·41848	·39882	·38122
0·70	0·67780	0·59693	0·54836	0·51217	0·48300	0·45852	0·43739	0·41875	0·40206
·71	·68467	·60729	·56117	·52689	·49932	·47623	·45632	·43877	·42306
·72	·69143	·61753	·57384	·54148	·51555	·49388	·47523	·45882	·44414
·73	·69810	·62764	·58637	·55594	·53166	·51143	·49409	·47886	·46525
·74	·70468	·63762	·59874	·57025	·54762	·52886	·51284	·49883	·48634
0·75	0·71116	0·64745	0·61096	0·58439	0·56342	0·54614	0·53147	0·51868	0·50735
·76	·71754	·65714	·62300	·59834	·57903	·56324	·54992	·53838	·52821
·77	·72382	·66669	·63487	·61210	·59443	·58012	·56816	·55788	·54888
·78	·73001	·67609	·64655	·62564	·60961	·59677	·58616	·57713	·56930
·79	·73610	·68534	·65804	·63897	·62454	·61316	·60388	·59609	·58943
0·80	0·74210	0·69443	0·66934	0·65206	0·63922	0·62926	0·62130	0·61474	0·60922
·81	·74800	·70337	·68043	·66492	·65362	·64506	·63839	·63303	·62864
·82	·75381	·71215	·69131	·67752	·66773	·66054	·65512	·65092	·64763
·83	·75952	·72078	·70198	·68986	·68154	·67567	·67147	·66840	·66617
·84	·76514	·72924	·71243	·70193	·69503	·69045	·68742	·68544	·68422
0·85	0·77067	0·73754	0·72266	0·71373	0·70821	0·70486	0·70295	0·70201	0·70175
·86	·77610	·74568	·73266	·72525	·72105	·71888	·71805	·71810	·71875
·87	·78144	·75366	·74244	·73649	·73355	·73252	·73270	·73368	·73519
·88	·78669	·76147	·75199	·74744	·74571	·74575	·74689	·74874	·75105
·89	·79184	·76912	·76132	·75810	·75752	·75857	·76061	·76327	·76632
0·90	0·79691	0·77660	0·77040	0·76847	0·76898	0·77097	0·77386	0·77727	0·78099
·91	·80188	·78392	·77926	·77854	·78008	·78296	·78663	·79072	·79505
·92	·80677	·79107	·78789	·78832	·79082	·79453	·79891	·80363	·80851
·93	·81156	·79806	·79628	·79780	·80120	·80567	·81071	·81599	·82135
·94	·81627	·80489	·80444	·80698	·81122	·81640	·82202	·82780	·83359
0·95	0·82089	0·81155	0·81237	0·81587	0·82089	0·82671	0·83286	0·83907	0·84522
·96	·82542	·81805	·82007	·82447	·83021	·83660	·84321	·84981	·85626
·97	·82987	·82439	·82754	·83277	·83917	·84608	·85310	·86001	·86671
·98	·83423	·83057	·83478	·84079	·84778	·85515	·86252	·86970	·87659
·99	·83851	·83659	·84180	·84852	·85605	·86382	·87149	·87887	·88591
1·00	0·84270	0·84245	0·84860	0·85597	0·86398	0·87210	0·88001	0·88755	0·89468

Table 1 (cont.). *The probability integral of the mean deviation (m) in normal samples of n observations. (Population standard deviation as unit)*

m \ n	2	3	4	5	6	7	8	9	10
1·00	0·84270	0·84245	0·84860	0·85597	0·86398	0·87210	0·88001	0·88755	0·89468
·01	·84681	·84816	·85518	·86315	·87158	·87999	·88809	·89574	·90292
·02	·85084	·85371	·86154	·87005	·87885	·88751	·89575	·90347	·91065
·03	·85478	·85911	·86768	·87668	·88581	·89465	·90299	·91074	·91788
·04	·85865	·86436	·87362	·88305	·89245	·90144	·90984	·91756	·92464
1·05	0·86244	0·86946	0·87935	0·88916	0·89878	0·90788	0·91629	0·92397	0·93095
·06	·86614	·87442	·88487	·89502	·90482	·91398	·92237	·92997	·93681
·07	·86977	·87922	·89020	·90063	·91057	·91976	·92809	·93557	·94227
·08	·87333	·88389	·89533	·90600	·91604	·92521	·93346	·94081	·94733
·09	·87680	·88841	·90027	·91114	·92123	·93037	·93850	·94568	·95202
1·10	0·88020	0·89280	0·90502	0·91605	0·92616	0·93523	0·94322	0·95022	0·95635
·11	·88353	·89705	·90959	·92074	·93084	·93980	·94764	·95444	·96034
·12	·88679	·90117	·91398	·92521	·93527	·94411	·95176	·95835	·96403
·13	·88997	·90516	·91820	·92947	·93947	·94815	·95561	·96198	·96741
·14	·89308	·90901	·92224	·93353	·94344	·95195	·95920	·96533	·97052
1·15	0·89612	0·91275	0·92613	0·93740	0·94718	0·95551	0·96253	0·96843	0·97337
·16	·89910	·91635	·92985	·94108	·95072	·95885	·96564	·97128	·97598
·17	·90200	·91984	·93341	·94458	·95406	·96197	·96851	·97391	·97836
·18	·90484	·92321	·93682	·94790	·95720	·96488	·97118	·97633	·98054
·19	·90761	·92647	·94009	·95105	·96016	·96760	·97365	·97855	·98252
1·20	0·91031	0·92961	0·94321	0·95404	0·96294	0·97014	0·97594	0·98058	0·98432
·21	·91296	·93264	·94620	·95687	·96555	·97250	·97805	·98245	·98595
·22	·91553	·93556	·94905	·95955	·96800	·97470	·97999	·98415	·98743
·23	·91805	·93838	·95177	·96209	·97030	·97675	·98178	·98571	·98877
·24	·92051	·94109	·95437	·96449	·97246	·97865	·98343	·98713	·98998
1·25	0·92290	0·94371	0·95685	0·96676	0·97448	0·98041	0·98495	0·98842	0·99107
·26	·92524	·94623	·95921	·96890	·97637	·98204	·98634	·98959	·99206
·27	·92751	·94865	·96146	·97092	·97813	·98355	·98761	·99066	·99294
·28	·92973	·95098	·96360	·97283	·97978	·98495	·98878	·99162	·99373
·29	·93190	·95322	·96564	·97462	·98132	·98624	·98985	·99250	·99445
1·30	0·93401	0·9553[illegible]	0·96758	0·97631	0·98275	0·98743	0·99082	0·99329	0·99508
·31	·93606	·9574[illegible]	·96942	·97790	·98408	·98852	·99172	·99401	·99566
·32	·93807	·9594[illegible]	·97117	·97940	·98533	·98953	·99253	·99465	·99616
·33	·94002	·9613[illegible]	·97283	·98081	·98648	·99046	·99327	·99523	·99662
·34	·94191	·9631[illegible]	·97441	·98213	·98755	·99132	·99394	·99576	·99702
1·35	0·94376	0·9644[illegible]	0·97591	0·98337	0·98855	0·99210	0·99455	0·99623	0·99738
·36	·94556	·96[illegible]2	·97733	·98453	·98947	·99282	·99510	·99665	·99770
·37	·94731	·96[illegible]4	·97867	·98562	·99033	·99348	·99560	·99703	·99798
·38	·94902	·96[illegible]9	·97995	·98664	·99112	·99409	·99606	·99736	·99823
·39	·95067	·97[illegible]7	·98115	·98759	·99186	·99464	·99647	·99767	·99845
1·40	0·95229	0·97[illegible]0	0·98229	0·98849	0·99254	0·99515	0·99684	0·99794	0·99865
·41	·95385	·97[illegible]5	·98337	·98932	·99316	·99561	·99717	·99818	·99882
·42	·95538	·97[illegible]4	·98439	·99010	·99374	·99603	·99748	·99839	·99897
·43	·95686	·97[illegible]9	·98535	·99083	·99427	·99641	·99775	·99858	·99910
·44	·95830	·97[illegible]7	·98626	·99151	·99477	·99676	·99799	·99875	·99922
1·45	0·95970	0·97[illegible]1	0·98712	0·99215	0·99522	0·99708	0·99821	0·99890	0·99932
·46	·96105	·97[illegible]9	·98793	·99274	·99564	·99737	·99841	·99904	·99941
·47	·96237	·98[illegible]3	·98869	·99329	·99602	·99763	·99859	·99915	·99949
·48	·96365	·98[illegible]1	·98941	·99380	·99637	·99786	·99874	·99926	·99956
·49	·96490	·98[illegible]6	·99009	·99427	·99669	·99808	·99889	·99935	·99962
1·50	0·96611	0·98385	0·99073	0·99472	0·99699	0·99828	0·99901	0·99943	0·99967

Table 1 (cont.). *The probability integral of the mean deviation (m) in normal samples of n observations. (Population standard deviation as unit)*

m \ n	2	3	4	5	6	7	8	9	10
1·50	0·96611	0·98385	0·99073	0·99472	0·99699	0·99828	0·99901	0·99943	0·99967
·51	·96728	·98471	·99133	·99513	·99726	·99845	·99913	·99950	·99972
·52	·96841	·98552	·99190	·99551	·99751	·99861	·99923	·99957	·99976
·53	·96952	·98630	·99243	·99586	·99774	·99875	·99932	·99962	·99979
·54	·97059	·98704	·99293	·99619	·99795	·99888	·99940	·99967	·99982
1·55	0·97162	0·98774	0·99340	0·99650	0·99814	0·99900	0·99947	0·99971	0·99984
·56	·97263	·98841	·99384	·99678	·99831	·99911	·99953	·99975	·99987
·57	·97360	·98905	·99425	·99704	·99847	·99920	·99959	·99978	·99989
·58	·97455	·98966	·99464	·99728	·99861	·99929	·99964	·99981	·99990
·59	·97546	·99023	·99500	·99750	·99875	·99936	·99968	·99984	·99992
1·60	0·97635	0·99078	0·99534	0·99771	0·99887	0·99943	0·99972	0·99986	0·99993
·61	·97721	·99130	·99566	·99790	·99898	·99949	·99975	·99988	·99994
·62	·97804	·99179	·99596	·99808	·99908	·99955	·99978	·99990	·99995
·63	·97884	·99226	·99624	·99824	·99917	·99960	·99981	·99991	·99996
·64	·97962	·99270	·99650	·99839	·99925	·99964	·99983	·99992	·99996
1·65	0·98038	0·99312	0·99675	0·99852	0·99932	0·99968	0·99986	0·99993	0·99997
·66	·98110	·99352	·99698	·99865	·99939	·99972	·99987	·99994	·99997
·67	·98181	·99390	·99719	·99877	·99945	·99975	·99989	·99995	·99998
·68	·98249	·99425	·99739	·99887	·99951	·99978	·99990	·99996	·99998
·69	·98315	·99459	·99758	·99897	·99956	·99980	·99992	·99996	·99998
1·70	0·98379	0·99491	0·99775	0·99906	0·99960	0·99983	0·99993	0·99997	0·99999
·71	·98441	·99521	·99791	·99915	·99964	·99985	·99994	·99997	·99999
·72	·98500	·99550	·99806	·99922	·99968	·99986	·99994	·99998	·99999
·73	·98558	·99577	·99820	·99929	·99971	·99988	·99995	·99998	·99999
·74	·98614	·99603	·99834	·99936	·99974	·99989	·99996	·99998	·99999
1·75	0·98667	0·99627	0·99846	0·99941	0·99977	0·99991	0·99996	0·99999	0·99999
·76	·98719	·99650	·99857	·99947	·99979	·99992	·99997	·99999	·99999
·77	·98769	·99671	·99868	·99952	·99982	·99993	·99997	·99999	·99999
·78	·98817	·99691	·99878	·99956	·99983	·99993	·99998	·99999	1·00000
·79	·98864	·99710	·99887	·99960	·99985	·99994	·99998	·99999	
1·80	0·98909	0·99729	0·99895	0·99964	0·99987	0·99995	0·99998	0·99999	
·81	·98952	·99746	·99903	·99967	·99988	·99995	·99998	·99999	
·82	·98994	·99762	·99911	·99970	·99990	·99996	·99999	1·00000	
·83	·99035	·99777	·99917	·99973	·99991	·99996	·99999		
·84	·99074	·99791	·99924	·99976	·99992	·99997	·99999		
1·85	0·99111	0·99804	0·99930	0·99978	0·99993	0·99997	0·99999		
·86	·99147	·99817	·99935	·99980	·99993	·99997	·99999		
·87	·99182	·99829	·99940	·99982	·99994	·99998	·99999		
·88	·99216	·99840	·99945	·99984	·99995	·99998	·99999		
·89	·99248	·99850	·99949	·99986	·99995	·99998	1·00000		
1·90	0·99279	0·99860	0·99953	0·99987	0·99996	0·99998			
·91	·99309	·99869	·99957	·99988	·99996	·99999			
·92	·99338	·99878	·99960	·99989	·99997	·99999			
·93	·99366	·99886	·99963	·99990	·99997	·99999			
·94	·99392	·99894	·99966	·99991	·99998	·99999			
1·95	0·99418	0·99901	0·99969	0·99992	0·99998	0·99999			
·96	·99443	·99907	·99972	·99993	·99998	·99999			
·97	·99466	·99914	·99974	·99993	·99998	1·00000			
·98	·99489	·99920	·99976	·99994	·99998				
·99	·99511	·99925	·99978	·99994	·99999				
2·00	0·99532	0·99930	0·99980	0·99995	0·99999				

T... (cont.). *The probability integral of the mean deviation (m) in normal samples of n observations. (Population standard deviation as unit)*

m \ n	2	3	4	5	6
·00	0·99532	0·99930	0·99980	0·99995	0·99999
·01	·99552	·99935	·99981	·99995	·99999
·02	·99572	·99940	·99983	·99996	·99999
·03	·99591	·99944	·99984	·99996	·99999
·04	·99609	·99948	·99986	·99997	·99999
2·05	0·99626	0·99951	0·99987	0·99997	0·99999
·06	·99642	·99955	·99988	·99998	·99999
·07	·99658	·99958	·99989	·99998	1·00000
·08	·99673	·99961	·99990	·99998	
·09	·99688	·99964	·99991	·99998	
2·10	0·99702	0·99966	0·99992	0·99998	
·11	·99716	·99969	·99992	·99999	
·12	·99728	·99971	·99993	·99999	
·13	·99741	·99973	·99994	·99999	
·14	·99753	·99975	·99994	·99999	
2·15	0·99764	0·99977	0·99995	0·99999	
·16	·99775	·99979	·99995	·99999	
·17	·99785	·99980	·99996	·99999	
·18	·99795	·99982	·99996	·99999	
·19	·99805	·99983	·99996	·99999	
2·20	0·99814	0·99984	0·99997	0·99999	
·21	·99822	·99985	·99997	1·00000	
·21	·99831	·99987	·99997		
·23	·99839	·99988	·99997		
·24	·99846	·99989	·99998		
2·25	0·99854	0·99989	0·99998		
·26	·99861	·99990	·99998		
·27	·99867	·99991	·99998		
·28	·99874	·99992	·99998		
·29	·99880	·99992	·99999		
2·30	0·99886	0·99993	0·99999		
·31	·99891	·99993	·99999		
·32	·99897	·99994	·99999		
·33	·99902	·99994	·99999		
·34	·99906	·99995	·99999		
2·35	0·99911	0·99995	0·99999		
·36	·99916	·99996	·99999		
·37	·99920	·99996	·99999		
·38	·99924	·99996	·99999		
·39	·99928	·99997	·99999		
2·40	0·99931	0·99997	1·00000		
·41	·99935	·99997			
·42	·99938	·99997			
·43	·99941	·99998			
·44	·99944	·99998			
2·45	0·99947	0·99998			
·46	·99950	·99998			
·47	·99952	·99998			
·48	·99955	·99998			
·49	·99957	·99999			
2·50	0·99959	0·99999			

m \ n	2	3
2·50	0·99959	0·99999
·51	·99961	·99999
·52	·99964	·99999
·53	·99965	·99999
·54	·99967	·99999
2·55	0·99969	0·99999
·56	·99971	·99999
·57	·99972	·99999
·58	·99974	·99999
59	·99975	·99999
2·60	0·99976	0·99999
·61	·99978	·99999
·62	·99979	1·00000
·63	·99980	
·64	·99981	
2·65	0·99982	
·66	·99983	
·67	·99984	
·68	·99985	
·69	·99986	
2·70	0·99987	
·71	·99987	
·72	·99988	
·73	·99989	
·74	·99989	
2·75	0·99990	
·76	·99991	
·77	·99991	
·78	·99992	
·79	·99992	
2·80	0·99992	
·81	·99993	
·82	·99993	
·83	·99994	
·84	·99994	
2·85	0·99994	
·86	·99995	
·87	·99995	
·88	·99995	
·89	·99996	
2·90	0·99996	
·91	·99996	
·92	·99996	
·93	·99997	
·94	·99997	
2·95	0·99997	
·96	·99997	
·97	·99997	
·98	·99998	
·99	·99998	
3·00	0·99998*	

* 0·99999 is reached for $m = 3·07$; 1·00000 is reached for $m = 3·23$.

Table 2. *Percentage points of the probability integral of the mean deviation (m), with the population standard deviation as unit*

(*a*) Lower percentage points

Size of sample n	0·1 %	0·5 %	1·0 %	2·5 %	5·0 %	10·0 %
2	0·001	0·004	0·009	0·022	0·044	0·089
3	0·022	0·052	0·073	0·116	0·166	0·238
4	0·066	0·114	0·145	0·199	0·254	0·328
5	0·112	0·170	0·203	0·260	0·315	0·386
6	0·153	0·215	0·250	0·306	0·360	0·428
7	0·190	0·252	0·287	0·342	0·394	0·459
8	0·220	0·283	0·318	0·372	0·422	0·484
9	0·247	0·310	0·344	0·396	0·445	0·504
10	0·271	0·333	0·366	0·417	0·464	0·521
Normal approximation:						
10	0·171	0·269	0·316	0·386	0·445	0·514

(*b*) Upper percentage points

Size of sample n	10·0 %	5·0 %	2·5 %	1·0 %	0·5 %	0·1 %
2	1·163	1·386	1·585	1·821	1·985	2·327
3	1·117	1·276	1·417	1·586	1·703	1·949
4	1·089	1·224	1·344	1·489	1·590	1·806
5	1·069	1·187	1·292	1·419	1·507	1·693
6	1·052	1·158	1·253	1·366	1·445	1·613
7	1·038	1·135	1·222	1·325	1·397	1·550
8	1·026	1·116	1·196	1·292	1·358	1·499
9	1·016	1·100	1·175	1·264	1·326	1·457
10	1·007	1·086	1·156	1·240	1·299	1·422
Normal approximation:						
10	1·000	1·069	1·128	1·198	1·245	1·342

BOOK REVIEW

The Advanced Theory of Statistics. Vol. I. By MAURICE G. KENDALL. London: Charles Griffin and Co. Ltd., 1943. Pp. 457. Price 42*s*.

It is difficult to review the present volume without knowing precisely how the author will deal with the topics reserved for its promised successor. For although Mr Kendall expresses the hope that this first instalment can profitably be read before the publication of the second, it is clear; nevertheless, that the two parts will be complementary and that full justice can only be done to this first part after the two volumes have been considered together.

Mr Kendall defines his objective as the provision of a systematic treatment of statistical theory as it exists at the present time. The work is encyclopaedic and will receive little criticism on the grounds of what is omitted. It is not an elementary book, the various topics being all carried to an advanced stage and at times requiring of the reader considerable mathematical powers. As with advanced theoretical work in most of the sciences, the practical problems which originally suggested the discussions have often receded well into the background. This is not mentioned as a criticism, for the sole value of scientific work does not necessarily lie in it being of immediate or even of ultimate practical importance. It is, however, proper to point out that Mr Kendall is here, in the main, content to present us with a picture of statistical theory as he finds it. The more controversial job of assessing the value of the different parts of the structure, whether from a purely practical or from an aesthetic viewpoint, he leaves to the reader. Within these self-imposed limitations he has scored a notable success.

The first six chapters deal in some detail with the properties of frequency distributions. Chapter 3, entitled 'Moments and Cumulants', is here particularly satisfying. It develops concisely the general relationships between the various families of power statistics. The simple presentation of the transformations giving moments in terms of cumulants and vice versa will be welcomed, and the listing of a large number of the resulting formulae will enhance the value of the book as a reference work. In addition, the familiar corrections for grouping, due to Sheppard, are derived, and the rather subtle distinctions between the conditions necessary for their application and the conditions under which the so-called average corrections for grouping may be applied are clearly drawn.

Chapter 4, entitled 'Characteristic Functions', begins with a proof of the Inversion Theorem which states that the characteristic function uniquely determines the distribution function. It then discusses at length various theorems connected with the limits of infinite sequences of distribution functions, and with the so-called Problem of Moments, i.e. the problem of specifying conditions under which the moments determine uniquely the distribution function. As in discussions of the convergence of infinite series in pure mathematics, the interest in the limits of sequences of distribution functions is almost exclusively theoretical, and this chapter should appeal to anyone who enjoys himself in this type of work. The more practical question of determining approximations with the aid of only a few moments is considered in later chapters.

Chapter 5 introduces the simpler distributions which are of central importance in statistical practice. Chapter 6 continues with a description of the Pearson system of curves and of the series developments of the Normal and Poisson distributions, associated with the names of Gram, Charlier and Edgeworth. One welcomes the fact that both the Gram-Charlier and the Edgeworth developments from the normal distribution are given, for although these series may be but rearrangements of each other, it is the order and grouping of the terms that are all important when any practical applications are intended. Indeed, although this point is well made, it is of such importance that, even at the risk of appearing to labour it, the addition of some further numerical illustrations might have been useful.

The next five chapters deal with Theories of Probability, with Sampling and with Sampling Distributions. On the first topic the author rightly does not dogmatize, judging that the rest of the subject appears to go forward in much the same way, whatever are one's basic concepts as to the meaning of probability. This determination not to take sides in the present book is extended also to the problem of induction, where he minimizes the too violent contrasts which have in recent years been drawn between Bayes's Theorem and the Principle of Maximum Likelihood. Far from being diametrically opposed, Mr Kendall observes that, if some account is taken of the limiting processes by which continuous distributions are defined, and if Bayes's Postulate is introduced in an appropriate manner, the two principles have a very strong resemblance.

The chapter devoted to the so-called 'exact' sampling distributions follows familiar lines. Perhaps of greater interest is Chapter 11, which is entitled 'Approximations to Sampling Distributions'. Here

the problems associated with the distribution of the k-statistics are given a very full treatment. It is, moreover, an authoritative treatment since Mr Kendall has himself contributed so much to the elucidation of the methods which are here required. The k-statistics, which were introduced by R. A. Fisher in 1928, have the property that their expected values in repeated samples are the cumulants of the populations sampled. In general the exact distributions of the k-statistics cannot be derived, but the cumulants of the k-distributions can be obtained by following out certain rules which were given by Fisher. Mr Kendall describes these rules and gives full proofs of their validity. A large number of the formulae derived from applying the rules in particular cases are quoted for reference. This is a very live subject, and one feels that there is still some scope for development and simplification of these methods, for it must still be admitted that their successful application requires a high degree of virtuosity. Aesthetically, however, they are far in advance of the heavy algebraic manipulation demanded by the earlier approach. The mathematician will perhaps find most in the chapter to satisfy his sense of what has come to be called elegance in such studies.

After this the reviewer found the remainder of the book dealing with the χ^2 distribution and with the Theory of Correlation something of an anti-climax. Most of all at this stage one feels the need to refer to the promised second volume, to which has been assigned the general theory of regression analysis. Correlation and regression are so closely allied that even a temporary separation is distressing.

While the above remarks may give some idea of the nature of the subjects discussed in this first volume, they do not do justice to the thoroughness with which the author has accomplished his task. It is by no means a book to skim through. Mr Kendall notes rather ruefully in his preface that statistical theory is essentially mathematical and suggests that it is not easy to keep the mathematics from getting on top. He says, however, that he intends his work to be one on statistics and not on statistical mathematics. The distinction is a fine one, and I must confess to some difficulty in appreciating Mr Kendall's point here. Readers who are familiar with Dr Aitken's small work entitled *Statistical Mathematics* may perhaps ask whether his book ought not, for the same reasons, to be termed one on statistics and not on statistical mathematics. However, by whatever name we speak of this subject, anyone who can follow a mathematical argument, and who has also at least some small experience of practical statistical problems, will find plenty to reward him in a study of Mr Kendall's book.

B. L. WELCH

(*All Rights reserved*)

BIOMETRIKA. Vol. XXXIII, Part III

CONTENTS

PAGE

On the use of matrices in certain population mathematics. By P. H. Leslie . . . . . 183—212

Errors in the routine daily measurement of the puerperal uterus. By C. Scott Russell . . 213—221

On a method of estimating frequencies. By J. B. S. Haldane . . . . . 222—225

The mathematics of a population composed of k stationary strata each recruited from the stratum below and supported at the lowest level by a uniform annual number of entrants. By H. L. Seal . . . . . 226—230

Moments of r and χ^2 for a fourfold table in the absence of association. By J. B. S. Haldane . 231—233

The use of χ^2 as a test of homogeneity in a $(n \times 2)$-fold table when expectations are small. By J. B. S. Haldane . . . . . 234—238

The treatment of ties in ranking problems. By M. G. Kendall . . . . . 239—251

The probability integral of the mean deviation:

Editorial note. By E. S. Pearson . . . . . 252—253

On the distribution of the estimate of mean deviation obtained from samples from a Gaussian population. By H. J. Godwin . . . . . 254—256

Appendix: Note on the calculation of the distribution of the estimate of mean deviation in normal samples. By H. O. Hartley . . . . . 257—258

Tables of the probability integral of the mean deviation in normal samples . . . . 259—265

Book Review:

M. G. Kendall's The advanced theory of statistics. Vol. I. By B. L. Welch . . . 266—267

PRINTED IN ENGLAND BY W. LEWIS, M.A., AT THE CAMBRIDGE UNIVERSITY PRESS

Reprinted by Photo-Lithography by Warren & Son Ltd., Winchester.

PAGE
183—212
213—221
222—225
226—230
231—233
234—238
239—251
252—253
254—256
257—258
259—265
266—267